© 2017 by Jonathan Lai. All rights reserved.

PROCESSES INVOLVED IN PROTEIN TRANSLATION

BY

JONATHAN LAI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Zaida Luthey-Schulten
Professor Martin Gruebele
Professor Taejkip Ha
Assistant Professor Thomas Kuhlman

Abstract

Protein translation—the process of converting genetic information into protein chains—is one of the oldest biological processes on planet Earth. Because of its age, all organisms, after the last universal common ancestor, use the same basic components—such as the transfer RNA (tRNA); aminoacyl-tRNA synthases; and the ribosome and its various elongation, initiation, and release factors—to carry out the translation process. It is the latter group, ribosomes and associated proteins, that physically make the peptide chains.

In the following thesis, we will explore the assembly and function of two components and their role in the protein synthesis. The thesis has been divided into several chapters describing (1-2) the assembly of the small subunit of the ribosome (SSU), and (3) free energy change and the mechanism of elongation factor Tu (EF-Tu). The function and evolutionary history of both components are intertwined as the SSU is responsible for decoding information encoded in messenger RNA (mRNA) and EF-Tu is responsible for delivering amino acid building blocks to the ribosome.

Acknowledgements

This work would not have been possible without the support of many people. First and foremost, I would like to thank my adviser, Zan Luthey-Schulten, who has helped me formulate strategies to tackle scientific problems and to ask pointed scientific questions.

I would like to thank my fellow lab members, both past and present, for their insight and advice. I enjoyed working with Ke Chen, John Eargle, Tyler Earnest, Zhaleh Ghaemi, Tyler Harpole, Yuhang Tang and Trang Thi Diem. Conversations with David Bianche, Marian Breuer, John A. Cole, Michael Hallock, Piyush Labhsetwar, Li Li, Yanxin Liu, Ryan McGreevy, Marcelo Cardoso Dos Reis Melo, Joseph R. Peterson, James C. Phillips, Elijah Roberts, and Seth Thor have been very rewarding. In particular, I would like to thank Ke Chen and Zhaleh Ghaemi for their guiding hand on several projects.

I would also like to thank my dissertation committee members Professor Martin Gruebele, Taejkip Ha and Thomas Kuhlman for providing me with insight and expertise in biophysics and cellular assembly.

Collaborations with other scientists from across the globe have helped me grow as a researcher. I would like to thank my collaborators Professors Sarah Woodson (Johns Hopkins), Sanjaya Abeysirigunawardena (Kent State), Hajin

Kim (University of Ulsan), Jennifer Morrell-Falvey (Oak Ridge National Lab), and Mitchel Doktycz (Oak Ridge National Lab).

I thank the NSF and DOE for funding me throughout my Ph.D studies and to the various computational centers around the country for providing me with the resources to conduct my studies.

Finally, I would like to thank my parents, sisters, friends, and dance friends/partners for all of their understanding and support. A special thanks goes to John C. Hadley, Tina Ho, Estelle Kao, Erica Lai, and Tiffany Lai for their spiritual and emotional support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Assembly of the 5′ domain of the small subunit of the ribosome[†]

1.1 Abstract

Using all-atom explicit solvent molecular dynamics simulations, we investigated the early structural intermediates of the 5′ domain of the 16S rRNA in *Escherichia coli (E. coli)* upon the removal of the primary binding r-proteins S4, S17, and S20, and the secondary binding r-protein S16. The removal of each r-proteins corresponded to the disappearance of subdomains with correlated dynamics. Correlation based network analysis of the MD trajectories showed that the different subdomains are connected via multiple pathways with high betweenness. These pathways cross at the internal loop of helix 17 (h17) in the five-way junction (5WJ). The structure of the internal loop is disrupted by the binding of S17 and rescued by the addition of S16, suggesting an important function of the secondary binding protein in biasing the rRNA folding landscape towards the native basis. Using structure-based Gō simulations, we investigated the folding of the lower four-way junction (4WJ) with h6, which is the primary binding site of S20 and the first to be

[†]Work includes previously published material and includes contributions from Ke Chen, John Eargle, Hajin Kim, Sanjaya Abeysirigunawardena, Sarah Woodson, Taejkip Ha, and Zan Luthey-Schulten. Published material referenced are as follows: Lai, et. al. [1], Chen, et. al. [2]

transcribed. The time course of folding of the 4WJ is consistent with the protection times observed in hydroxyl radical footprinting. Results from the all-atom simulations show that the fluctuations in the 5WJ are independent of the fluctuations in the 4WJ—suggesting that the subdomains fold independently and are stabilized by primary r-proteins.

1.2 Introduction

To decipher the complexities of ribosome assembly, one must simultaneously understand the cooperative coordination of ribosomal proteins (r-proteins) and the co-transcriptional folding of the ribosomal RNA (rRNA). Nomura *et. al* first addressed this process using *in vitro* reconstitution experiments of the *E. coli* ribosomal small subunit (SSU), which led to the first assembly map showing the hierarchical dependency of the r-proteins upon association with the 16S rRNA [3]. Recent progress in biophysical approaches have added further details to this assembly pathway. Particularly, using time-resolved hydroxyl radical footprinting, Woodson and coworkers suggested that early assembly of the 30S subunit proceeds through multiple parallel pathways nucleated at different positions on the 16S rRNA [4, 5]. Furthermore, advances in pulse-chase quantitative mass spectrometry have yielded kinetic data for all r-proteins during *in vitro* assembly of the ribosomal SSU [6–8]. These findings agree on the 5′ to 3′ directionality during 30S ribosomal assembly and emphasize the need to consider both co-transcriptional folding of the rRNA and cooperative binding of the r-protein to study ribosomal

biogenesis *in vivo* [9].

Although the 5′ domain of the 16S rRNA has been shown to be capable of folding and establishing some of its final tertiary contacts without additional r-proteins [4,10], cryo-EM studies have indicated that well-resolved folding intermediates are only observed after the binding of 5′ primary binding r-proteins S4, S17, and S20 [8]. Prior to the binding of these proteins, the 5′ domain rRNA likely exists as an ensemble of disordered structures. Hydroxyl radical footprinting studies suggest that these structures may be associated with different combinations of bound r-proteins as each of the primary binding r-proteins appear to favor distinct intermediate conformations [11]. Despite significant experimental progress, many details of how the association of different r-proteins is cooperatively coupled with the folding of the 5′ domain remain unclear.

Application of high performance computing has the advantage of yielding a comprehensive view of the assembly process at high structural and energetic resolution. Molecular dynamics (MD) simulations, both all-atom and coarse-grained, have been successfully applied to examine functional motions of the intact ribosome [12–16]. As ribosomal assembly occurs over many different time scales, reduced representations have been used to study the stability of the 30S ribosomal subunit, providing insight into the correlations between the order of protein binding with electrostatic energy and changes in RMSD of the SSU [17–21]. To obtain a higher resolution structural description of the assembly intermediates, we started our investigation with all-atom molecular dynamics simulation of the five-way junction (5WJ:

h3, h4, h16-h18), which is the minimal binding site for the initiator binding protein S4 [2, 22]. Taking advantage of the hierarchical folding of the RNA, we removed S4 and all Mg^{2+} ions from the 5WJ structure to disrupt the tertiary contacts and explore the topology of its folding landscape [2, 23]. Using angles between pairs of helices as the reaction coordinate, we were able to identify three folding intermediates around the native state. These structural intermediates were consistent with findings in the accompanying FRET and SHAPE experiments, and the energetic relationships among them were resolved by monitoring the refolding of the 5WJ upon addition of Mg^{2+} ions and S4. Interactions of the intrinsically disordered N-terminus of S4 with helix h16 suggested a fly-casting mechanism first introduced by Wolynes and co-workers [24–26] to achieve efficient protein:RNA recognition during early assembly. The simultaneous folding of the 5WJ and binding of S4, investigated under the principle of minimal frustration [27], further confirmed this scenario [2].

In this present work, we used similar approaches to probe structural intermediates of the full 5′ domain rRNA coupled to the binding of different r-proteins. The 5′ domain under investigation is comprised of rRNA helices h3 to h18, three primary binding proteins S4, S17, and S20, and one secondary binding protein S16, taken from the crystal structure of the *E. coli* 30S subunit [28]. Long simulations on the naked 5′ domain rRNA with monovalent ions and with various combinations of the r-proteins were compared to a reference state of the 5′ domain with all four proteins and Mg^{2+} ions. Differential fluctuations in the rRNA calculated from these simulations suggested

two subdomain structures in addition to the previously studied 5WJ, each capable of folding independently upon association with the primary binding proteins S17 and S20. To investigate the main barriers to the formation of the binding site for S20, structure-based Gō simulations were carried out that revealed key tertiary contacts required for establishing the lower four-way junction (4WJ: h7-10). Further network analysis revealed multiple communication pathways connecting motions of the individual subdomains that cross at a critical internal loop in h17, one of the binding sites of S16. Finally, comparisons of the fluctuations in the motions of the 5′ domain of the 16S rRNA to those observed in the truncated system without the 5WJ indicate the behavior of each subdomain can be studied independently. Our study suggests that the folding of the 5′ domain rRNA and ribosomal assembly proceeds through formation of independent subdomains, which correlated well with the observed barriers in formation of the local and non-local helices.

1.3 Methods

1.3.1 Molecular models of the 5′ domain rRNA

All-atom models of the 5′ domain with and without the associated r-proteins were built using the crystal structure of the *E. coli* ribosomal SSU (PDB code: 2I2P) [28]. To compare to the previous footprinting experiments and assembly maps [8], nucleotides 21 to 562 (*E. coli* numbering) of the 16S rRNA and various combinations of ribosomal proteins S4, S16, S17, and S20 (Table 1.1, #1 to #9) were included in the models. In addition, two reduced

rRNA constructs were set up to study the folding of the S20 binding domain (primarily the 4WJ and helix h6) without proteins. One was constructed from nucleotides 47 to 239 to describe the 4WJ. The second (nucleotides 47 to 393) includes the helices h5-15, but omits the 5WJ.

Binding of S17 disrupted the key interactions in the h17 internal loop, while S16 restored them (see Results). In order to visualize this process, a representative conformation of the 5′ domain rRNA was chosen from the end of the simulations with S4, S17, and with S4, S20, respectively. S16 was then added to the partially unfolded 5′ domain, 8 Å away from its binding site in the crystal structure. Steric clashes between S16 and the rRNA were manually resolved by adjusting the proteins side-chains of A27, G30, R31, F32, and K76 in a flexible loop. In both cases, no modifications were made to the primary binding proteins (S4 and S17, or S4 and S20). The resulting structure was then subject to the minimization and thermalization protocol presented in the next section.

1.3.2 All-atom molecular dynamics simulations

Proteins and nucleic acids were parameterized with the CHARMM22 with CMAP corrections [29] and CHARMM27 [30] force fields, respectively. All systems, as summarized in Table 1.1, were prepared using the protocol established in Eargle *et al.* [31,32]. In the reference 5′ domain simulation with four r-proteins (Table 1.1, #1), 22 Mg^{2+} within 4.5 Å of the crystal structure (2I2P) were included in the system. To neutralize the system, an additional 59 Mg^{2+} and 392 Na^{+} ions were placed according to local electrostatic potential

by *Ionize* [33]. The Mg^{2+} and Na^{+} ions were placed 2.5 Å and 6 Å away from the solute respectively. Six water molecules were then explicitly coordinated around each Mg^{2+} ion to complete its hexahydrated solvation shell, except for the 8 Mg^{2+} ions in direct contact with the rRNA backbone. In all other simulations, neutralizing K^{+} ions were placed 6 Å away from the rRNA and r-proteins similarly by *Ionize*. Neutron scattering studies have shown that water is more tightly bound to RNA than protein [34]. As such, careful attention was paid in solvating our systems. The first five solvation layers were placed using the *Solvate* software [35], and then the *VMD* [36] solvate plugin was used to complete the water box with a minimum 17 Å buffer region on each side of the protein:RNA complex. The TIP3P water model was used. Finally, additional K^{+} and Cl^{-} ions were added to reach a physiological ion concentration of 150 mM, except for the simulations of the reference complex and the naked rRNA (Table 1.1, #2). The resulting all-atom systems range in size from 296,000 to 480,000 atoms. Further details describing each system are listed in Table 1.1.

MD simulations were performed using *NAMD 2.9* [37]. As the local solvent and ion density around the RNA molecules is different than around proteins [32,34], all prepared systems were minimized and equilibrated in a step-wise fashion. Minimization was carried out using the conjugate gradient method in *NAMD*, first with positional constraints on all heavy-atoms for 2,000 steps. Constraints were then released from the water molecules for 3,000 steps. Protein and nucleic acid side-chains, as well as the ions, were set free for the next 5,000 steps. Finally, all atoms were set free for the

last 20,000 steps of minimization. Thermalization was conducted using a temperature jump protocol with step-wise positional restraints to allow waters and ions to diffuse slowly into and pack against the RNA structure. The initial temperature was set to 100K, and ions and heavy atoms in the RNA and protein were harmonically restrained for 25 ps. Then, the temperature was raised to 200K, and ions and the backbone atoms were harmonically restrained for 25 ps. In the next step, the backbone atoms were harmonically restrained at the temperature of 250K for another 50 ps. Force constants for all harmonic restraints were set to 1 kcal·mol$^{-1}\cdot$Å^{-1}. Finally, the temperature was raised up to 300K and all atoms were freed for further equilibration.

Production runs were conducted under the NPT ensemble with pressure and temperature held at 1.01325 bar and 300 K using the Langevin thermostat (5 ps^{-1} damping coefficient) and Langevin barostat (200 fs piston period and 100 fs piston decay). Periodic boundary conditions were applied, and multiple time-stepping was used to calculate bonded interactions at 1 fs, van der Waals (vdW) interactions every 2 fs, and electrostatic forces every 4 fs. Particle mesh Ewald summation was used to evaluate electrostatic interactions, and the cutoff for the vdW force calculations were 12 Å with a switching distance of 10 Å. A total of 928 nanoseconds (ns) of all-atom MD simulation on the 5′ domain were reported in this paper.

1.3.3 Structure-based Gō model

Several structural based potentials were attempted to model the assembly of ribosome assembly. The first comes a simple LJ native contact potential

ID	Proteins included	Starting conf.	rRNA Residues	Total # atoms	Dimensions (Å)	T (ns)	# Neutralizing ions (K^+, Na^+)/ Mg^{2+}	Buffer region (Å)
1	S4, S16, S17, S20	crystal str.	21 to 562	340,000	179×131×136	100	392 / 81	20
2	-	crystal str.	21 to 562	370,000	140×184×137	156	541 / 0	24
3	S4	crystal str.	21 to 562	465,000	174×176×146	158	524 / 0	25
4	S4, S17	crystal str.	21 to 562	408,000	181×134×161	100	517 / 0	23
5	S4, S20	crystal str.	21 to 562	480,000	196×145×160	100	508 / 0	28
6	S4, S17, S20	crystal str.	21 to 562	480,000	161× 150×189	100	501 / 0	24
7	-	crystal str.	47 to 393	296,000	136×121×173	101	347 / 0	22
8	S4, S16, S17	100ns of #4	21 to 562	480,000	203×143×174	40	517 / 0	17
9	S4, S16, S20	100ns of #5	21 to 562	480,000	196×140×169	40	508 / 0	20
10	S20	Gō intermediate	47 to 239	301,690	137×174×125	33	167 / 5	21

Table 1.1: Summary of all-atom molecular dynamics simulations performed in this paper.

with bonded terms (including bonds, angles, dihedrals and impropers) inherited from the CHARMM force field. All non-bonded potentials (vdW and electrostatics) were replaced by the knowledge based Gō-potential. The Gō-potential is formulated with respect to a reference structure, which can either be the equilibrated native structure for the folding of RNA/protein or a target conformation that is important for the function of the biomolecules. For atom pairs closer than 4 Å within a molecular chain or pairs closer than 4.4 Å between the RNA and protein chains in the reference structure, the pair is defined as a native contact and is subject to a Lennard-Jones-style potential:

$$V_{native}(\sigma_{ij}^{native}, r_{ij}) = 4\epsilon^{native}\left[\left(\frac{\sigma_{ij}^{native}}{r_{ij}}\right)^{a} - \left(\frac{\sigma_{ij}^{native}}{r_{ij}}\right)^{b}\right] \tag{1.1}$$

where σ_{ij}^{native} is determined by the native pairwise distance, r_{ij}^{native}, between atoms i and j in the reference structure:

$$\sigma_{ij}^{native} = \left(\frac{b}{a}\right)^{\frac{1}{b-a}} r_{ij}^{native} \tag{1.2}$$

such that the potential reaches the minimum at σ_{ij}^{native}. r_{ij} is the instantaneous distance between atoms i and j in the simulation, ϵ^{native} specifies the depth of the potential, and the exponents – a, b – determine the shape of the potential. All non-native atom pairs with a distance longer than 4 Å in the reference structure experience a repulsive potential of the form:

$$V_{nonnative}(r_{ij}) = \epsilon^{nonnative} \left(\frac{\sigma^{nonnative}}{r_{ij}} \right)^{c} \tag{1.3}$$

where the non-native $\epsilon^{nonnative}$ scales the strength of the potential, $\sigma^{nonnative}$ determines where the potential equals $\epsilon^{nonnative}$, and the exponent, c, determines the long range decay.

The above described hybrid MD-Gō model is implemented within the newest version of NAMD2, and our implementation allows for the study of complex systems in which molecules with different properties are present at the same time. It is designed so that each molecule may have its own parameters set independently of the others. In the particular system that we are studying in this paper, we set $a = 12$ and $b = c = 6$ for the folding of both protein and nucleic acid. The native interaction strength ϵ^{native} is set to 0.1 kcal·mol^{-1} and 0.23 kcal·mol^{-1} for RNA and protein, respectively, to ensure that S4 and the 5WJ have similar transition temperatures. All contacts within protein and RNA, irrespective of being involved in sidechain-sidechain, sidechain-backbone, backbone-backbone, sugar-sugar, sugar-base, or base-base interactions, are treated equally. To facilitate binding initiated from large separations, protein-nucleic interaction potential is set with exponents $a = 8$ and $b = 4$ and $\epsilon^{native} = 0.15$ kcal·mol^{-1}. All non-native

parameters are set to $\epsilon^{nonnative} = 0.01$kcal·mol^{-1}, and $\sigma^{nonnative} = 2.5$ Å in this study.

1.3.4 Hybrid MD-Gō folding simulations

The hybrid MD-Gō simulations were prepared without water molecules or ions, and only the heavy atoms of the 5WJ and S4 protein were used. Two representative states of the partially unfolded 5WJ (discussed in Results and Discussion) were chosen from the all-atom MD simulations as the starting conformations for the simultaneous folding and binding simulations. Given the high degree of secondary structure stability of S4, especially in its globular C-terminal domain [38, 39], we chose a slightly unfolded structure with an overall RMSD of ∼5.2 Å in which the disordered S4 N-terminus has a RMSD of 7.2 Å and the globular S4 C-terminal domain has a RMSD of 4.5 Å. The unfolded S4 and 5WJ were placed ∼38 Å (center-of-mass distance) away from each other, such that the closest contacts between them are just under the non-bonded interaction cutoff (12 Å). A 2 fs time step was used for all hybrid MD-Gō simulations. Each simulation was run for 2,250,000 steps, such that if the 5WJ and S4 were not bound within this time, they were likely to diffuse away from each other. One hundred replicates starting from each chosen 5WJ conformation were performed with parameters introduced in the previous section, and statistics were generated.

1.3.5 Structure-based Gō model of the S20 binding domain

An all-atom structure-based model (*SMOG* version 1.2.1) [40–42] was applied to the S20 binding domain (h6 and 4WJ, nucleotides 47-239) to study its assembly process. All interactions were described by a cut-off contact map using the following default nucleic acid forcefield parameters [41]. Contacts were considered native if the distance between nucleic-nucleic residues was less than 4Å. Native contacts were described by a 6-12 Lennard-Jones (LJ) potential and subjected to a 15 Å cutoff. The ratio between the total contact to dihedral energy was set to 2.0. All dihedrals were parameterized to have the same strength. Excluded volumes were described by the repulsive part of the LJ potential, with ϵ_{NC} set to 0.01 kcal·mol^{-1}, and σ_{NC} set to 2.5 Å.

A fully extended S20 binding domain was generated in SMOG using high temperature unfolding and used as the initial conformation. A total of 27 replicate simulations were performed to investigate structural intermediates and conformational changes along the folding pathway. The system was first subjected to 20,000 steps of limited-memory BFGS minimization, and then allowed to refold at temperature well-below the predicted melting temperature (0.6 T_m $\approx$ 65 K). All simulations were integrated every 0.5 fs, and ran for 20,000,000 time steps each in *Gromacs 4.5.5* [43].

1.3.6 Analysis of the simulated trajectories

The root-mean-square fluctuation (RMSF) of the MD trajectories was calculated by first aligning the 5′ domain rRNA backbone to the crystal structure.

Once aligned, the RMSFs of the phosphorous atom for each nucleotide were calculated from the last 75 nanoseconds of each MD trajectory. To qualitatively determine the role of each r-protein binding, the difference between RMSF values between relevant simulations was calculated for each nucleotide. Large decreases in RMSF, with a magnitude greater than the standard deviation, indicated that the nucleotide were stabilized by the addition of the protein. Likewise, large increases in RMSF indicated that the nucleotide were destabilized by the addition of the protein.

Angles between RNA helices were calculated between helical axes of a pair of chosen helices. Helical axis was approximated from the third principal axes of inertia, and boundaries of the helices were determined based on the secondary structure diagram [44] (Figure 1.1). Three helices with complex secondary structures, h11, h17, and h18, were described by multiple helical vectors. Specifically, h11 was represented by three helical vectors: h11a (nucleotides 240 – 245, 283 – 288); h11b (nucleotides 275 – 279, 246 – 250); and h11c (nucleotides 251 – 274). Helix h17 was represented by two helical vectors: h17a (nucleotides 437 – 447, 488 – 494) and h17b (nucleotides 451 – 482). Two vectors were used to represent h18: h18a (nucleotides 505 – 509, 521 – 528) that described the pseudo-knot region; and h18b (nucleotides 499 – 504, 510 – 520, 533 – 546). The angles were calculated between every pair of the helical vectors for each frame in the trajectory and translated into the range [0°, 90°]. The standard deviation for each angle over time was calculated for further conformational clustering.

Correlation based network analysis was performed similar to that pre-

sented in previous studies of protein and rRNA complexes [32, 45–47]. Each nucleotide in the 5′ domain was described using two nodes: one located on the phosphorous atom, and the other located on the N9 nitrogen in adenosine and guanosine or the N1 nitrogen in cytidine and uridine. Network edges were defined between a pair of nodes if the closest distance between any heavy atoms represented by the corresponding nodes was less than 4.5 Å for at least 75% of the trajectory. Close contacts between nearest neighbors in sequence were excluded. Networks were generated from trajectories sampled every 20 ps. The weight of an edge, w_{ij}, is calculated from the Pearson correlation, C_{ij}, between nodes i and j from the simulation such that $w_{ij} = -log(|C_{ij}|)$. The length of a path D_{ij} connecting nodes i and j is the sum of edge weights between nodes (k, l) along the path: $D_{ij} = \sum_{k,l} w_{kl}$. The shortest path is the most correlated series of edges between two nodes and is calculated using the Floyd-Warshall algorithm. The betweenness of an edge is defined as the number of shortest paths going through that edge for all pairs of nodes. Edges with the largest betweenness values are the main avenues of communication and only the top 7% are mapped onto the 3D structure of the 5′ domain using the Network View plugin in VMD [47].

Fraction of native contacts (Q) measures the similarity between a unfolded structure and its native conformation that ranges between 0 (no similarity) to 1 (identical) [48, 49]. The value of Q is given as the ratio of native contacts formed to the total number of native contacts. For our purposes, a native contact is defined among pairs of atoms not in the same residue but whose separation is less than 4 Å. A native contact is considered

formed if the measured distance differs from the native distance by less than 20%. Native contacts were calculated with respect to the crystal structure. $Q_{secondary}$ and $Q_{tertiary}$ measured the Q for contacts within a given helix and between helices, respectively. A structure is considered folded once a frame's Q score reaches the time-averaged Q calculated over the last fifth of each trajectory. Q scores were calculated using the Libbiokit package released as part of MultiSeq [49] in VMD.

1.4 Results and Discussion

1.4.1 Differential role of S17 and S20 in early 5′ domain folding

Three of the six primary binding proteins for the ribosomal SSU, S4, S17 and S20, bind to the 5′ domain. Time resolved pulse chase mass-spectrometry and electron microscopy revealed that the initiator protein S4 fully binds to the SSU within one minute [8, 50]. The large number of positive charges (+17) that S4 carries, as well as the local contacts S4 establishes with the rRNA junction formed from h3, h4, h16, h17, and h18, provide some clues for the early association of S4. On the other hand, the low number of positive charges of S17(+6) and the global interactions that S20 makes with multiple helices (h6, h7, h8, h9, h11, h13, h14) may give rise to the delay for their binding to reach equilibrium (Figure 1.1).

To investigate how these primary binding r-proteins initiate and modu-

late the folding of the 5′ domain rRNA, we adopted a hierarchical simulation scheme in which we started with the naked 5′ domain rRNA and systematically added r-proteins to the RNA molecule in an ordered fashion according to the assembly map [8]. Six simulations were generated for this purpose: 5′ domain rRNA with no protein, with S4, with S4 and S17, with S4 and S20, with S4, S17 and S20, and with S4, S16, S17 and S20. All simulations started from the native conformation of the 5′ domain complex, and the global effect of each r-protein were extracted by comparing structural fluctuations in the rRNA explored in different simulations.

Without any proteins, the 5′ domain shows large fluctuations at both the 5′ and 3′ leader sequences of h3, as well as helices in the S20 binding domain (h6, h9, h10), S17 binding domain (h11, h12), and in the 5WJ (h16-h18) (Figure 1.2a). These helices are located on the peripheral area of the 5′ domain and are expected to fluctuate on the nanosecond to microsecond timescale. A totally different pattern in the nucleotide fluctuation was seen upon addition of S4 which interacted with only the 5WJ. First of all, the disappearance of peaks in the RMSF of h16, h17, and h18 (Figure 1.2b) indicated that S4 binding not only stabilized individual nucleotides, but also eliminated large conformational change in the 5WJ. This observation is consistent with our previous study showing that binding of S4 efficiently shifts the equilibrium of the 5WJ from the extended and misfolded states to its native conformation [2]. Second, fluctuations in other helices—such as h6, h9 to h12—were generally reduced. However, a similar shape in the RMSF curve suggests that the arrangement of these helices are not directly

modulated by the binding of S4. Interestingly, opposite to the behavior of most of the helices, h8 and h14 experienced a large movement upon binding of S4 (Figure 1.2b). In the crystal structure, these two helices interact with each other through an A-minor motif made in their end loops, and they together form bridge B8 on the interface with the large subunit [28].

S20, whose three-helix bundle spans a length of ~ 55 Å, makes contacts with seven different helices in the 5′ domain, and, therefore, the combination of S4 and S20 is expected to exert a more global effect on the rRNA folding than S4 alone. Indeed, RMSF values in the simulation of the 5′ domain with both S4 and S20 showed a global stabilization throughout the 5′ domain compared to that with S4 alone (Figure 1.2c). It is also worth noting that motions of h12 were greatly reduced, and the entire helix h11 was stabilized. Because h11 and h12 comprise the binding site for S17, their modulated dynamics by S20 strongly indicates correlated behaviors of S17 and S20 in the folding of the 5′ domain rRNA. Surprisingly, when S17 was included in the simulation, in addition to S4 and S20, all helices except h8, h11, h14 and h18 were destabilized (Figure 1.2d). Helices h8, h11, and h14 all make direct contacts with S20. This mutual effect implies that binding of S17 and S20 likely apply competing conformational restraints on the 5′ domain rRNA and may favor distinct intermediate states.

1.4.2 Binding of r-proteins give rise to subdomains in the 5′ domain

Changes in single-nucleotide fluctuations upon binding of S4 illustrated the protein's relatively localized effect on stabilizing mainly the 5WJ, which has been studied in detail previously [2]. On the other hand, the binding of S20 induced much more global stabilization of the 5′ domain, while the addition of S17 destabilized the 5′ domain with respect to the crystal structure. In order to identify the intermediate conformations favored by S17 and S20, it is necessary to deconvolute changes in the secondary and tertiary structure. Angles between pairs of helices naturally describe the organization of RNA helices and are, therefore, chosen to delineate changes in the tertiary structure.

Five helices, h5, h6a, h7, h13, and h15, displayed small angle fluctuations with respect to each other in all simulations (Figure 1.3a). Although h7, h13, h15 make direct contacts with r-proteins (Figure 1.1), these helices show backbone RMSD values less than 5 Å for over 150 ns with neither r-proteins nor Mg^{2+} present, suggesting that they form the "core" of the 5′ domain (data not shown). Interestingly, three out of the five (h5, h6a, h7) are non-local helices of which the two strands are separated in sequence, indicating possible late formation of this core subdomain.

As expected, S4 binding stabilized the core and 5WJ by restricting the relative motion between h16 and h18, consistent with previous findings obtained from both experiments and simulations [2] (Figure 1.3a). Helices

composing the 5WJ move concertedly after S4 is bound, while the 5WJ as a whole still fluctuates independently with respect to the other helices (Figure 1.3b,c). Furthermore, S4 also shows a long range effect that reduces motion of h6 and h8 with respect to the core subdomain.

Inclusion of S17 or S20 in addition to S4 in the simulation results in greatly reduced fluctuations in the remaining helices of the 5′ domain (Figure 1.3d). These helices are situated around the binding sites of S17 and S20 and are structurally separated by more than 40 Å along the core subdomain helix h7 (Figure 1.3e,f). Therefore, they are grouped into two subdomains for subsequent discussion: 1) the S17 binding domain (h11 and h12), and 2) the S20 binding domain (h6-10). In the simulation with S4 and S20, not only do the helical fluctuations within the subdomains decrease, but so do the relative motions between subdomains, indicating that S20 strongly favors the native conformation of the 5′ domain rRNA (Figure 1.3d,f). When combined with S4, S17 displays a weaker capacity for organizing tertiary structures than does S20 (Figure 1.3d). Although local helical motions are reduced, motions of the S17 binding domain and the S20 binding domain with respect to the 5WJ remain (Figure 1.3d,e).

To confirm that the subdomains behave independently, a reduced system, containing helices h5 to h15, was created to mimick the 5′ domain without the 5WJ (Table 1.1, #7). After simulating the system for over 100 ns, the reduced system showed similar helical motions, confirming that the subdomains behave independently from each other (Figure 1.6).

It is interesting to note that the heat maps for the simulations involving

the three primary binding r-proteins resemble the one calculated for the simulation involving only S4 and S17 (Figure 1.3d,g top halves). Similarly, the heat map for the simulation involving all four proteins (S4, S16, S17, and S20) resembles the one involving only S4 and S20 (Figure 1.3d,g bottom halves). The resemblance further confirms that S20 locks the 5′ domain in a native-like conformation, while binding of S17 biases the 5′ domain towards a non-native intermediate state even with the presence of S20. This intermediate arises from the relative motions between subdomains rather than orientations of any particular helices.

1.4.3 S16 partially restores S17-induced distortion in h17 internal loop

Observations from our MD simulations suggested that S17 increased the fluctuations between the 5WJ and S20 binding domain, while S16 restricted the relative motions between the two subdomains (Figure 1.3d,g). To probe how S17 and S16 directly affect the folding of the rRNA, correlation-based network analysis was used to identify communication pathways between these subdomains [45–47, 51].

Using the naked 5′ domain simulation as a guide, a weighted network was constructed from correlations of long-lived pair-wise contacts between nucleotides. The dynamic network of the naked 5′ domain rRNA revealed continuous pathways connecting the 5WJ to S20 binding domain through helices h15 and h17 (Figure 1.4a). The pathway connecting the two helices passes from A374 in h15 through an internal loop in h17 (nucleotides A451

and A452). Since helix h17 extends across the length of the 5′ domain, the conformation of the internal loop fixes the position of the 5WJ and S20 binding domain.

In order to see if different sets of bound r-proteins lead to different subdomain packings, the conformation of the internal loop was monitored. The relative distances between the center of mass of the internal loop nucleotides (448 – 455) and A374 with respect to the crystal structure were recorded. Two separate cases emerged. With the addition of S4, S16, or S20, the relative distances between internal loop backbone did not change much compared to the crystal structure (Figure 1.4b). With S17 bound, however, the 5′ half of the internal loop (nucleotides 448 – 451) moves farther away (≈ 6Å) from nucleotide A374, while the 3′ half (nucleotides 453 – 455) moves closer (≈ 4Å) (Figure 1.4b). Reintroducing S16 to the end of the S4 + S17 simulation appears to rescue the internal loop conformation. In 40 ns, S16 was able to move towards its binding site and push nucleotides G453 – G455 closer towards the crystal structure. As expected, adding S16 to the end of the S4 + S20 simulation had minimal effect on the internal loop conformation (Figure 1.4c).

1.4.4 Simulating assembly of the S20 binding domain

Being a secondary binding protein, S16 stabilizes the emerging subdomains once they are formed. In our previous studies, the tertiary contacts in the 5WJ were shown both computationally and experimentally to form upon binding of S4 [2]. Given that the fluctuations in the 5WJ in this study appear to

be independent of the other parts of the 5′ domain, formation of the S17 binding domain and S20 binding domain should only depend on S17 and S20, respectively. Considering the 2–3 orders of magnitudes difference between the *in vivo* RNA elongation rate and the zipping rate of RNA helices, we assume that formation of RNA secondary structure occurs concurrently with the transcription of the next helix [52, 53].

Examination of the secondary structure diagram, in the context of the 5′ to 3′ directionality of translation, suggests that the first pair of helices to be completely transcribed — and hence fold — is h6 and h8 in the S20 binding domain. In the crystal structure, helices h6 and h8 form tertiary contacts that join the 4WJ to the core subdomain in the 5′ domain. They, together with h7,9,10,11,13, and h14 make contacts with S20. Similar to the primary binding r-protein S4, S20 possesses a stable C-terminal domain that binds, in this case, tightly to its main binding site (h7 and h10). On the other hand, the disordered N-terminal domain of S20 only becomes structured upon binding to the rRNA [21, 54], suggesting a similar fly-casting mechanism in the protein:RNA recognition and binding process.

To study the assembly of the S20 binding domain, SMOG [40–42] was used to examine its folding starting from a fully extended conformation. A total of 27 simulations were generated to investigate the structural intermediates and energetic barriers on the folding landscape of the S20 binding domain. The fraction of native contacts within ($Q_{secondary}$) and between ($Q_{tertiary}$) helices was used to identify conformational changes in the secondary and tertiary structure along the folding pathway.

Without the help of S20, 14 out of 27 successful folding events of the S20 binding domain were observed. It appeared that all observed folding events proceeded through three steps: 1) folding of individual helices; 2) formation of the 4WJ (h7-h10); and 3) packing of h6 against the 4WJ.

Because the 5′ and 3′ strands of the local helices (h6 and h8-h10) were sequentially juxtaposed, they started to fold immediately after the simulations began. Intrahelical contacts ($Q_{secondary}$) within these helices formed in short bursts (Figure 1.5a-d). The fact that these helices began to fold at roughly the same time, and that the average time to helix formation increased with the number of base pairs in the helix, suggested that the local helices formed independently to each other. These observations enabled us to compare the base pair formation rate in our Gō simulation to that observed in optical tweezer experiments [55], which provided an estimation of the timescale in the Gō simulations (see Supporting Information). According to such approximation, helices h6, h8, h9, and h10 were measured to form within 24, 12.5, 6.5, and 8.5 μs respectively (Figure S2).

The non-local helix h7 differs from the other local helices in that its 5′ and 3′ strands are sequentially separated by ∼90 nucleotides. Furthermore, it contains a helix-internal loop-helix motif that divides this long helix into two co-axially stacked stems, which are referred to as h7a (nucleotides G122 – C136, G227 – U239) and h7b (nucleotides G138 – G142, C221 – C225). Both h7b and h7a started to fold after the formation of the h8-h10 local helices (Figure 1.5e,f). During folding of the local helices, h9 and h10 were oriented nearly co-linearly while h8 was oriented perpendicularly, so that

the helices adopted a "T-like" shape. This"T-like" shape was stabilized by tertiary interactions near the junction, and gave rise to a transient structural intermediate represented by $Q_{tertiary} \approx 0.5$ (Figure 1.5g). The h8-h10 T-like shape accelerated folding of h7, which in turn promoted the formation of the 4WJ ($Q_{tertiary} \approx 0.75$, Figure 1.5g).

In the last step, h6 was expected to establish tertiary interactions primarily with h8 in order to pack against the 4WJ into the native conformation. However, 12 single stranded nucleotides separating h6 from the 4WJ greatly increased the searching time for the correct docking orientation that placed h6 at a 110° angle with respect to h8 (Figure 1.5h). Before the successful docking, the helices underwent multiple conversions between a collapsed but mis-oriented state (radius of gyration $R_G \sim 35$ Å) and an extended state ($R_G \sim 55$ Å) until the native contacts between h6 and h8 were finally formed ($R_G \sim 33$ Å). Once the center of mass (CoM) distance of the contact region between h6 and h8 moved within 30 Å, the angle between h7 and h6 becomes restricted near 60° (Figure 1.5i). Early formation of the 4WJ sterically blocked h6 from approaching h8, or caused the formation of nonnative contacts, leading to the unsuccessful folding of the S20 binding domain in the remaining ∼50% of the simulated trajectories. Therefore, it is expected that formation of the native interactions between h6 and h8 might take several milliseconds to seconds, consistent with the slow hydroxyl radical protection of h6 and h8 over different Mg^{2+} concentrations [4].

Finally, we examined the possibility that binding of S20 might accelerate the docking between h6 and h8. We started an all-atom MD simulation

from a late structural intermediate where the 4WJ was fully formed and h6 was within 30 Å (CoM) from h8. A fully folded S20 was introduced to this intermediate state, with its C-terminal domain close to the 4WJ. In 33 ns, the C-terminal domain of S20 quickly bound to its binding site, while the exposed positively-charged residues on the N-terminal domain of S20 were able to promote localization between h6 and h8. The results suggested that a mechanism similar to S4 could be adopted by S20 for the final formation of the S20 binding domain (Figure 1.9).

1.5 Conclusion

Simultaneous folding of rRNA and binding of r-proteins to the 5′ domain produce a series of structural intermediates critical to the assembly of the SSU. In this paper, the assembly was studied using explicit solvent all-atom MD simulations. Analysis of the variations in helical angles of the 16S rRNA led to the identification of four dynamical subdomains: the core, the 5WJ, S20 binding domain (4WJ and h6) and S17 binding domain (h11 and h12). While helices in the core remained stable without the presence of the r-proteins, tertiary contacts in the other subdomains are only formed after the primary binding proteins S4, S17, and S20 bind. Network analysis based on correlations from the MD simulations of the naked rRNA showed multiple communication pathways connecting these subdomains and highlights how the binding of S17 intermittently destabilizes the protein:RNA complex by twisting the internal loop in h17. Binding of the secondary protein S16

appears to partially counteract the effects of S17 on the internal loop. In addition, a combination of all-atom and structure-based Gō simulations were used to study the unfolding and folding of the S20 binding domain. This work provides molecular details on how S20 stabilizes key tertiary contacts in the lower helical junction that are consistent with time dependent protection studies. The main barriers to the formation of the lower junction are the docking of the 3′ and 5′ strands of the non-local h7 followed by re-orientation of the helices in the four-way junction to allow tertiary contacts between h6 and h8 to be established. The all-atom simulations of the naked rRNA in the full and truncated 5′ domains further reveal that the motions of the 5WJ are independent of the lower helical junctions so that the assembly process can be considered to take place through the formation of independent subdomains.

1.6 Supporting Information

1.6.1 5′ domain residues involved in betweenness pathways

S4	S16	S17	S20
26	42	121	60
27	43	126	61
28	**44**	**127**	**62**
400	**108**	**128**	**63**
401	110	**129**	**102**
402	111	**130**	**103**
403	**112**	233	**104**
404	**134**	234	**105**
405	**135**	235	**106**
406	**136**	236	**107**
407	137	237	**108**
408	**138**	238	**131**
409	**226**	252	**132**
410	**227**	253	**133**
411	**228**	254	**175**
412	**229**	255	176
413	**230**	256	177
418	**231**	**264**	178
419	**308**	**265**	184
425	**309**	**266**	185
426	**310**	**267**	186
427	311	272	187
428	**325**	273	191
429	**326**	274	192
430	**373**	275	193

S4	S16	S17	S20
436	**374**	**276**	**194**
437	375	**277**	**195**
438	376	**278**	**196**
439	377	**279**	**222**
440	378	**280**	223
489	**389**	**301**	258
490	**390**	**302**	259
491	**391**		260
495	**392**		261
499	**393**		**262**
507	**449**		**263**
508	**450**		**264**
509	**451**		322
510	**452**		323
511	**453**		**324**
512	**473**		**325**
540	**474**		327
541	**483**		329
542	**484**		**331**
543	**486**		**332**
544			**333**
545			**350**
546			**351**
547			

Table 1.2: List of residues in the 5′ domain within 5 Å of a r-protein. Residues in bold indicate that the residues is a part of the high betweenness pathways connecting different subdomains.

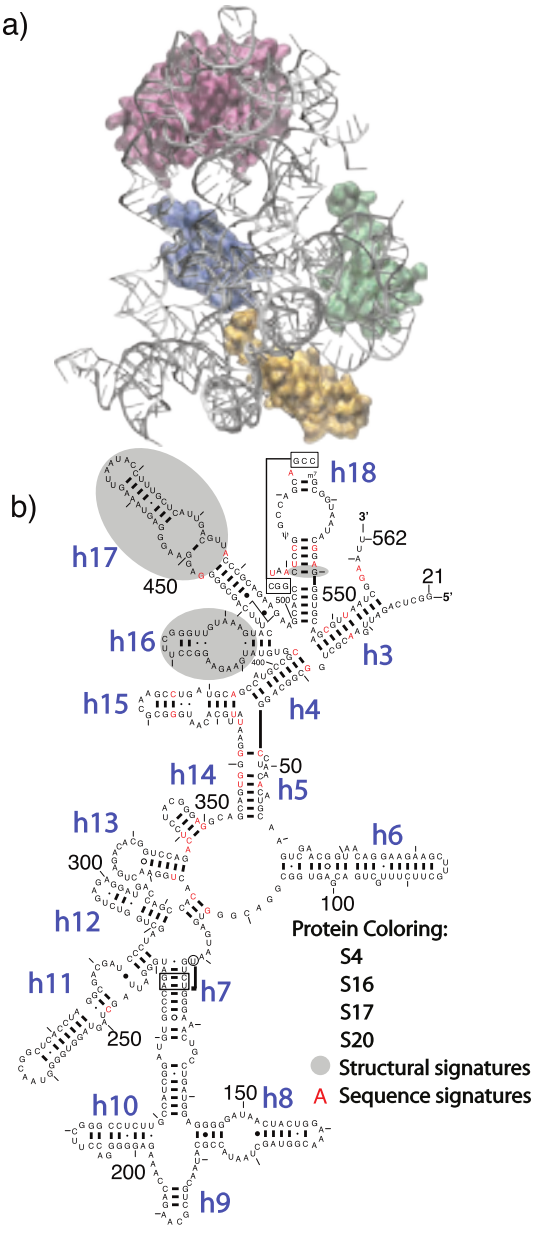

Figure 1.1: Protein:RNA contacts in the 5′ domain. (a) Protein:RNA interactions in the crystal structure (2I2P [28]. R-proteins are colored as follows: S4 (purple); S16 (blue); S17 (green); and S20 (yellow). (b) Binding sites of the r-proteins overlaided onto the 5′ domain secondary structure map adapted from the Comparative RNA Website (CRW) [56]. Contacts are defined to be within a 5 Å cutoff based on the crystal structure for each 5′ domain r-proteins. Sequence and structural signatures are highlighted on the diagram in red typeface and gray shading, respectively [57].

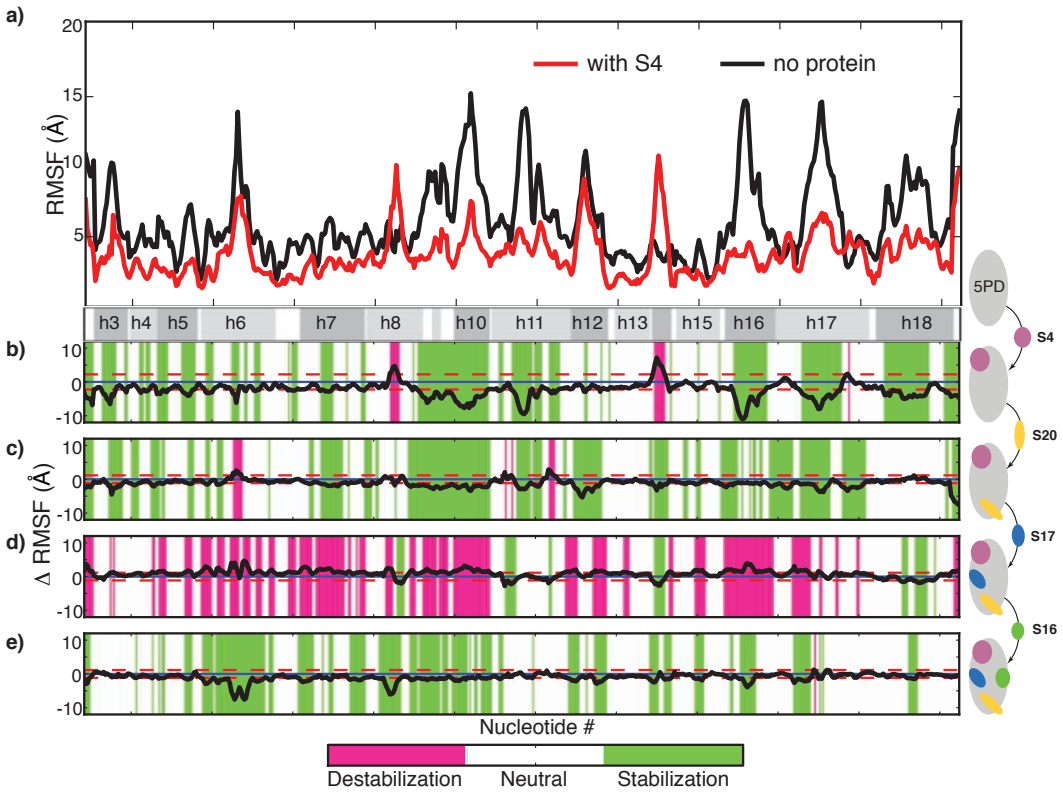

Figure 1.2: Global influences of r-protein binding on the dynamics of the 5′ domain rRNA. a) RMSF in the 5′ domain rRNA without any proteins and with S4 bound. Difference in RMSF between simulations: b) without and with S4; c) with S4 and with S4 + S20; d) with S4 + S20 and with S4 + S17 + S20; e) with S4 + S17 + S20 and with S4 + S16 + S17 + S20. The dashed red line indicates the standard deviation in the RMSF difference. Nucleotides stabilized by the corresponding protein are colored green, while ones that are destabilized are colored by purple (see Methods).

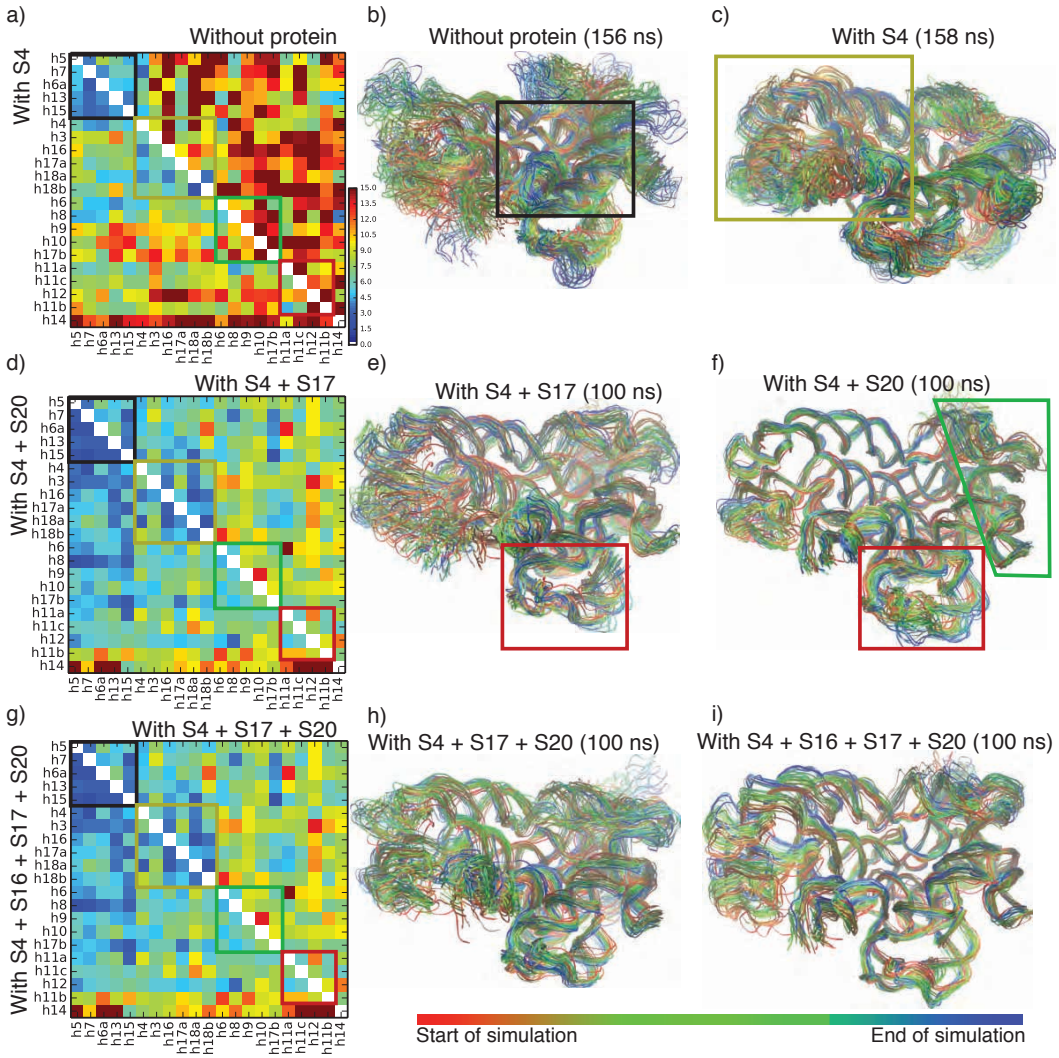

Figure 1.3: Formation of dynamical subdomains. Fluctuations of pairwise angles and time traces of the 5′ domain structure are shown for simulations: a-c) without protein and with S4, d-f) with S4 + S17 and S4 + S20, and g-i) with S4 + S17 + S20 and S4 + S16 + S17 + S20. Helices in the heat map are reordered for clarity. Boxes highlight groups of helices that form subdomains in the 5′ domain. The "core" subdomain (black) appears without any proteins while the 5WJ (tan) appears upon binding of S4. The S20 binding (green) and S17 binding (red) subdomains only appear after binding of S20 and S17. Time traces of MD trajectories are aligned by the core subdomain, and colored by time steps, with red representing the start of the simulation and blue showing the end.

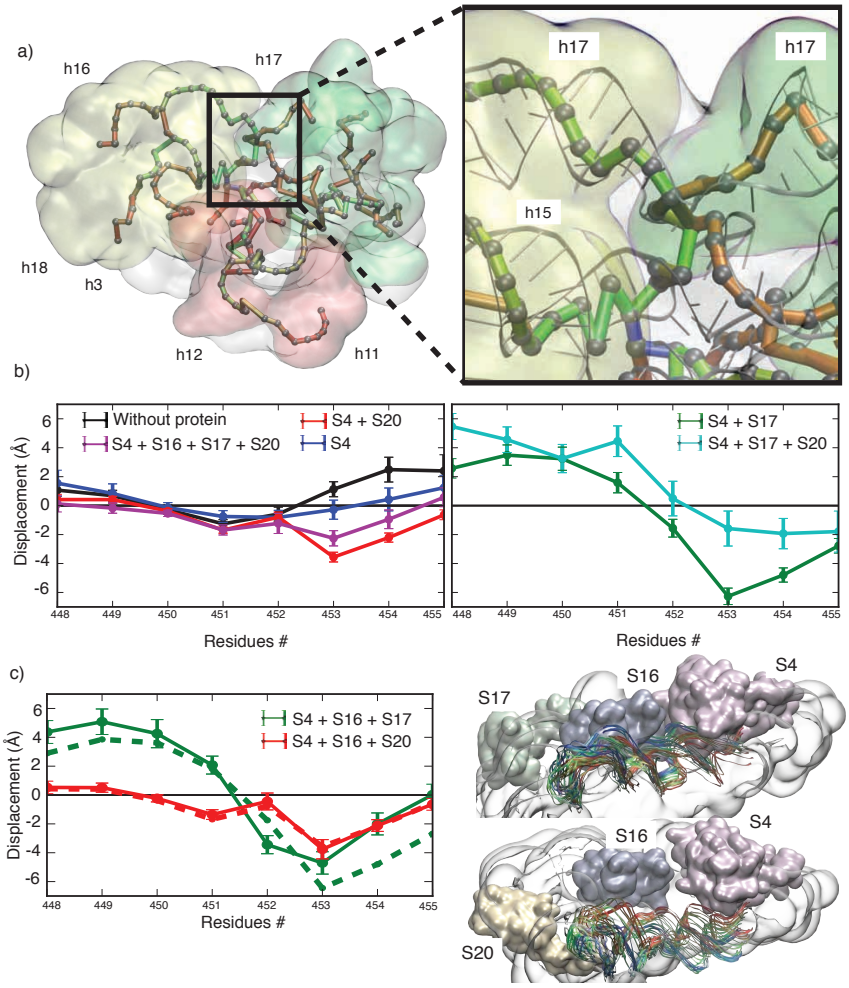

Figure 1.4: Conformational switch in the h17 internal loop. a) Correlation-based network calculated from the 5′ domain simulation without proteins (Table 1.1, #2). Edges with the top 7% of betweenness are shown and colored according to the betweenness values from red to blue. Helices, shown in quick surf, have been colored based on the 5WJ, S17, and S20 binding domains in Figure 1.3. b) Relative center of mass displacement of the nucleotides in the internal loop backbone (A448 to G455) to A374 in h15. Displacement calculated with respect to the crystal structure. Solid lines show the time average from 50 to 100 ns for each trajectory. Vertical bars show standard deviation in the displacement. S17 has the largest effect on the h17 internal loop. c) Addition of S16 in rescue simulations restore the internal loop. The initial coordinates were taken from the last frame of the S4 + S17 and S4 + S20 simulations. Dashed lines are taken from the corresponding simulations in panel b. Changes in internal loop structure over time shown in insert; again, color denotes 40ns time trace (red to blue).

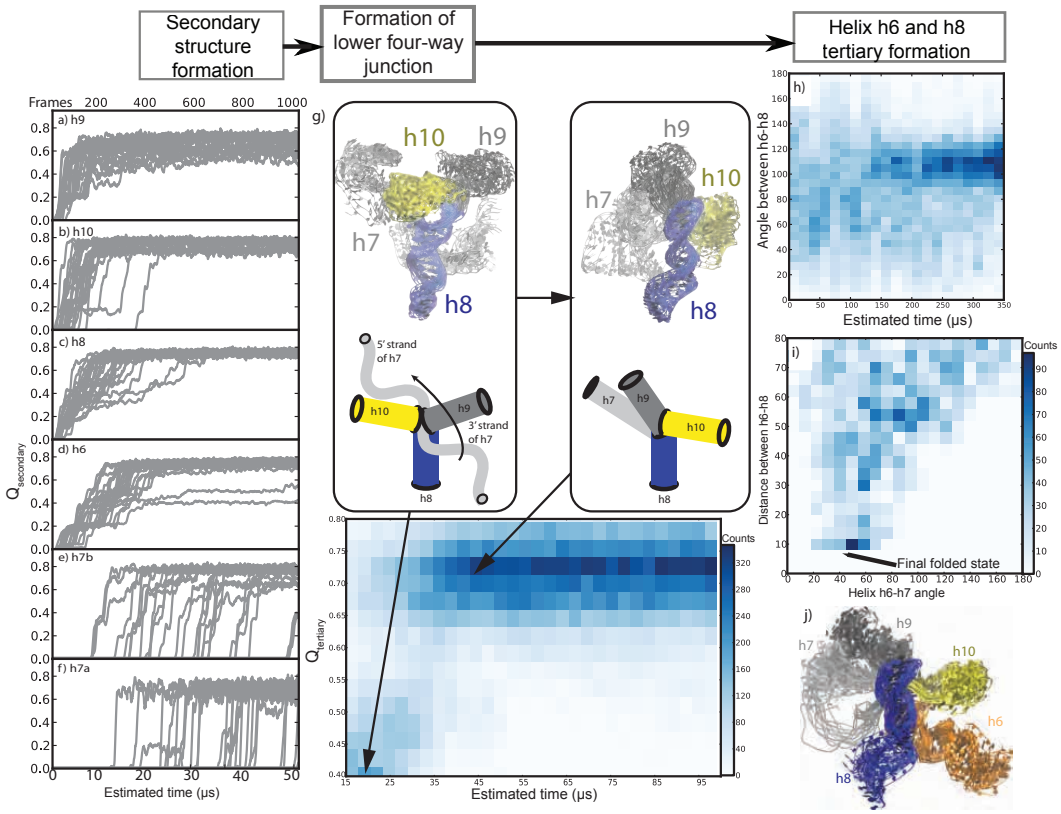

Figure 1.5: Timeline for the secondary and tertiary structure in the S20 binding domain. a-f) Running average of $Q_{secondary}$ time traces for helices h9, h10, h8, h7a, and h7b respectively. Time traces are taken from the first 100 μs of all 27 replicates. The running average is calculated using a 1 μs window. Traces in red greatly differed from the other runs. g) $Q_{tertiary}$ time trace for native contacts between helices h7b, h8a, h9, and h10 for all trajectories. Yellow line shows the average formation time of h6. Green line shows the average formation time of h7b. Color indicates number of frames for a given time and Q score. Snapshots of the trajectory are oriented such that helix h8 is always pointing downwards and helix h7 is always pointing into the plane. h-j) Shows the 14 folded trajectories where the h6-h8 tertiary contact was formed. h) Angle time traces between helices h6,h8 after h7 formation. i) Angle of approach between helices h6,h7 is fixed as h6 approaches h8. Heatmap shows restriction of the h6,h7 helical angle as the average contact distance between h6 and h8 shrinks. j) Helices h7-h10 in the folded state.

1.6.2 Calibrating a physical time for Gō simulations

Because a structure-based model was used, the simulation timesteps need to be calibrated to a physical timestep. This was accomplished by comparing the rate of nucleic acid zippering in the *SMOG* simulations as well as *in vitro* data. Recent studies suggest that base pairs (bp) zip at the rate of ~ 0.9 μs/bp [**?**,55]. For the SMOG simulations, base pair formation over time was calculated for each helix. Linear regression was used to estimate the helix zipping rate. Local helices had an average zipping rate of 7 ps/bp while the nonlocal helix h7 had a zipping rate of 14 ps/bp. Upon further analysis, each of the helices zipped in short bursts, with short segments zipping rates varying from 1 to 4.5 ps/bp, followed by several μs of pausing. In the case of the h7, the pausing lasted for tens of μs. Because of the variation in the pausing time, only a rough estimate for zipping can be calculated; thus, an average zipping rate of 4.5 ps/bp was assumed for all of the helices. By equating the simulated and real zipping rates, 5 ps of simulation time equals 1μs of physical-time. We use this to approximate the time required to form secondary and tertiary contacts in S20 shown in Figure 1.7.

1.6.3 Role of S20 in accelerating h6-h8 formation

The role of S20 in accelerating h6-h8 tertiary contact formation was studied using an all-atom NAMD simulation. As the starting conformation, an intermediate conformation was taken from the Gō simulations, where the closest contact between tips of helix h6 and h8 is more than 12 Å away from

each other. The 4WJ from this conformation was aligned to the X-ray crystal structure, allowing the S20 r-protein to be directly transferred to the new simulation from the reference PDB file. The few steric clashes incurred from this transfer were manually resolved by moving conflicting sidechain or backbone in S20 using *VMD*. The clashes only occurred in the N-terminus of S20. Because the N-terminus of S20 only becomes structured upon binding to the 5′ domain, any of these manual resolutions should have minimal effect on the S20 dynamics [54].

The starting conformation is shown in Figure 1.9. After alignment, the system was solvated, equilibrated, and simulated using the protocols listed in the methods section of the main document. During the 30 nanoseconds of simulation, the closest contact distance between the tip of h6, h8, and S20 were recorded; three lysines on S20 (K4, K7, K8) and sidechain of asparagine (N2) in S20 are responsible for tightly binding to h6 and h8 and they formed most of the closest contacts from S20 to the helices. Once bound to S20, the tips of helix h6 and h8 are stably brought together (Figure 1.9).

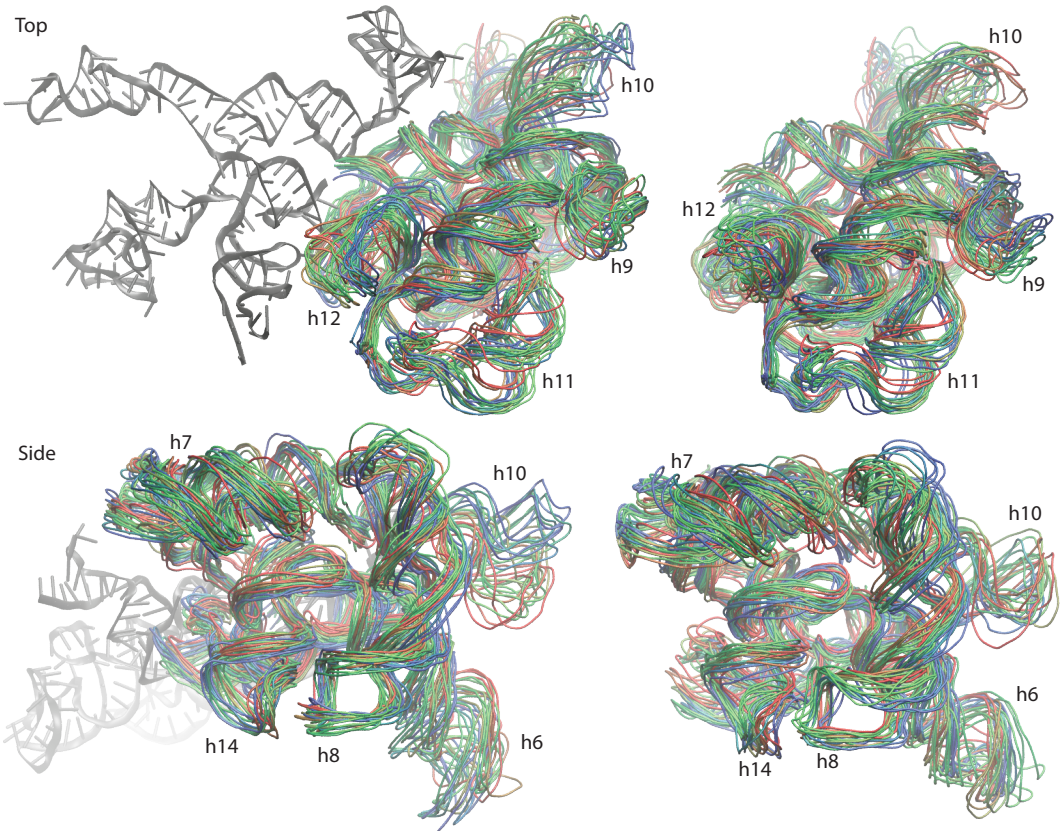

Figure 1.6: Time traces of the 5′ domain with and without (left and right columns respectively) the 5WJ (helix h5-h15). Without the 5WJ, the 5′ domain and the reduced system show similar fluctuations over the same time scale in all of the helices. This similarity suggests that the S17 and S20 binding domains are dynamically independent from the 5WJ. Traces are color coded from red to blue based on time (over the course of 100 ns).

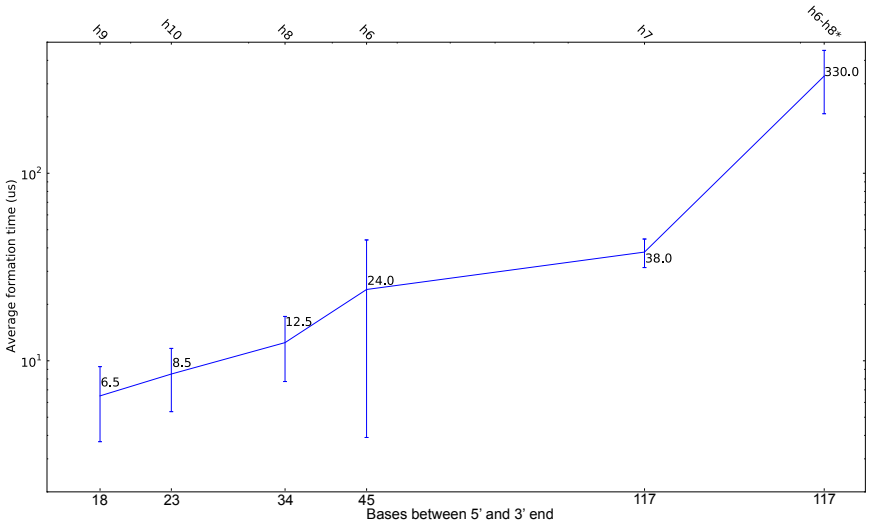

Figure 1.7: Estimated time needed to form secondary and tertiary structure. Helices are sorted by the length of sequence between the 5′ and 3′ ends. Times are calculated from all 27 folding replicates with the exception of the tertiary contact h6-h8*; for the tertiary contact, only folded replicates were included (14 out of 27).

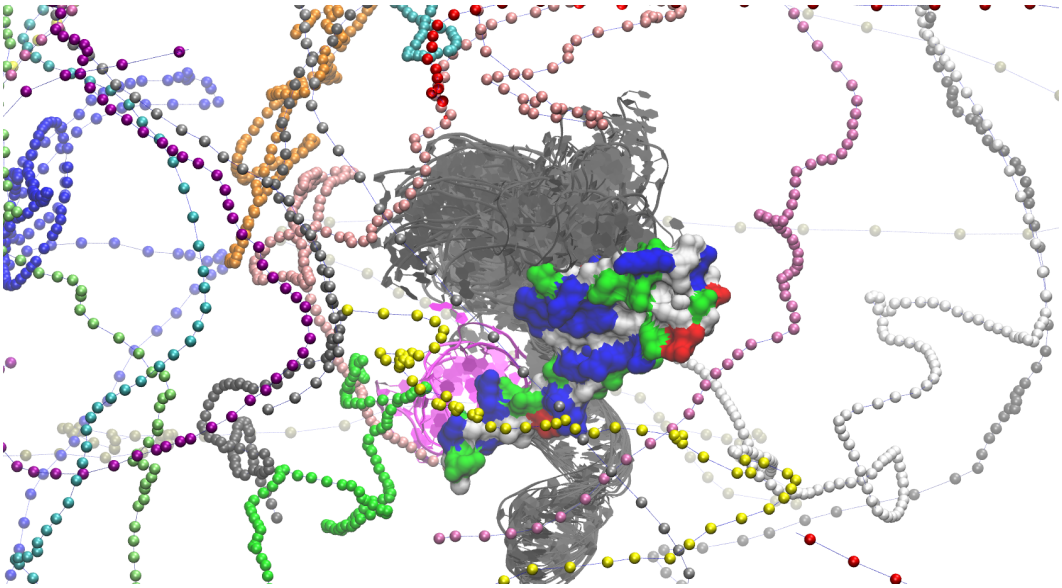

Figure 1.8: Folding runs aligned to the crystal structure by h8. Frames are synchronized around the formation of secondary structure in helices h7b-h10. Helices h7 through h10 have adopted their final tertiary conformation and provide a binding surface for S20. Dots show the center of mass for h6 from each of the folding trajectories as h6 attempts to dock to h8. The final binding site for helix h6 is shown in pink. Positively charged residues on S20 might help to extend the capture radius of h6 and accelerate the S20 binding domain formation 1.9.

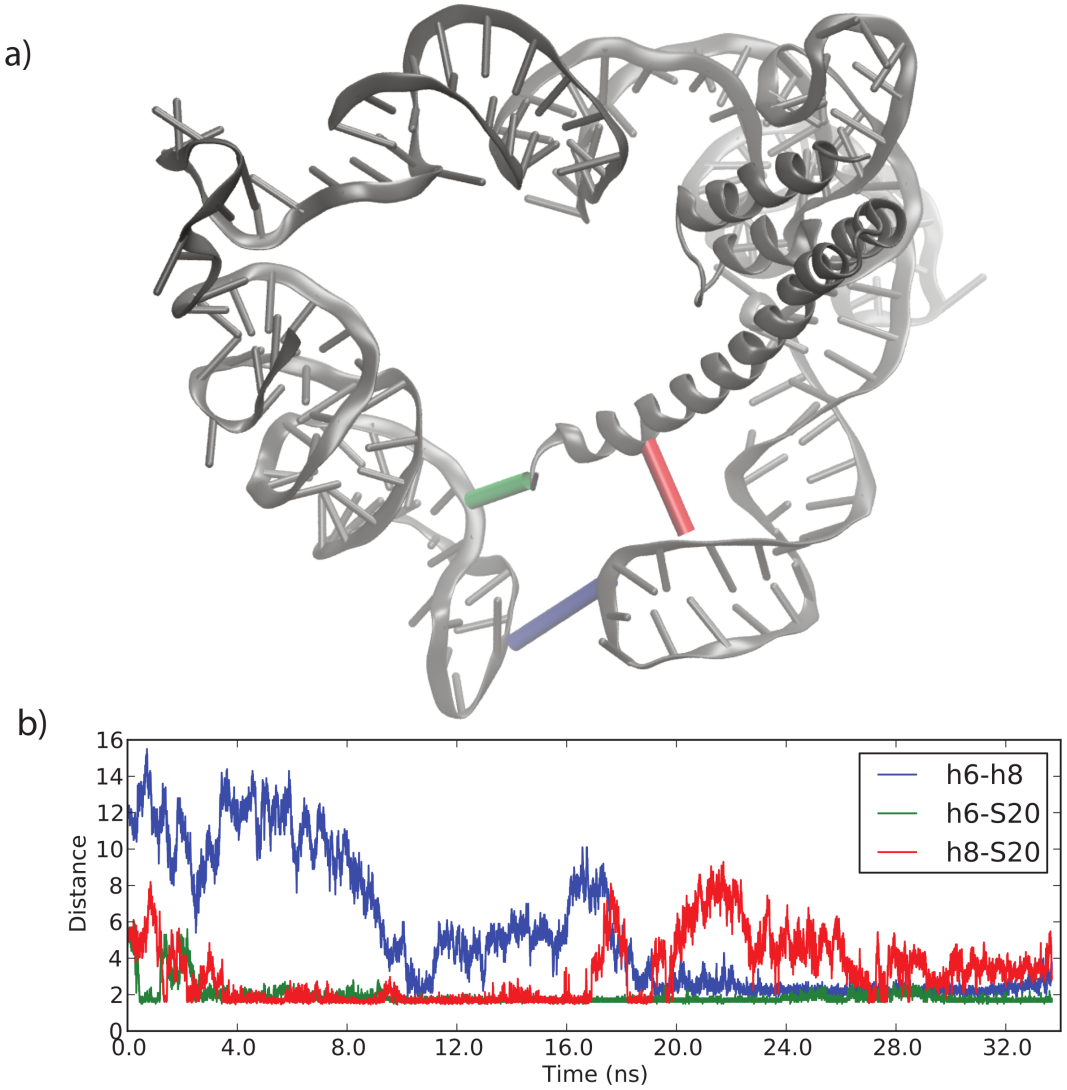

Figure 1.9: Closest contact distance between the r-protein S20 and the tips of helices taken from an all-atom simulation. Top figure shows the distances measured between helices. Bottom figure shows how the distances change over time. Tips are given as follows: (h6: nucleotides 76 to 92), h8 (h8: nucleotides 158 to 165), and S20 (S20: amino acids 2 to 20). Color describes distances between helix h6 and h8 (blue), h6 and S20 (green), and h8 and S20 (red).

Chapter 2

Assembly of the small subunit[†]

2.1 Abstract

Central to all life is the assembly of the ribosome: a coordinated process involving the hierarchical association of approximately 54 proteins to the RNAs forming the small and large ribosomal subunits. The process is further complicated by effects arising from the intracellular environment such as molecular crowders and the location of ribosomal operons within the cell. We report on our progress on the construction of a whole-cell model of ribosome biogenesis in *E. coli*. Kinetic models of small subunit reconstitution *in vitro* at the level of individual protein/rRNA interactions are developed for two temperature regimes. The model at low temperatures predicts the existence of a novel $5' \rightarrow 3' \rightarrow$ central assembly pathway, which we investigate further using molecular dynamics. The high temperature model is incorporated into a model of *in vivo* ribosome biogenesis, including the transcription and translation of ribosomal protein and RNA. This whole-cell model predicts the localization of early assembly intermediates to the nucleoid region and

[†]Work includes previously published material and includes contributions from Ke Chen, Tyler Earnest, and Zan Luthey-Schulten. Published material referenced are as follows: Earnest, et. al. [58], Abeysirigunawardena, et. al. [59]

reproduces the known assembly timescales for the small subunit with no modifications made to the embedded *in vitro* assembly network.

2.2 Methods

Atomic models of the assembly intermediates are built using the crystal structure of the *E. coli* ribosomal SSU (PDB: 2I2P) [28]. Proteins and nucleic acids are parameterized with the CHARMM36 [60, 61] force fields. All systems are prepared as follows:

Systems were neutralized by placing sodium ions according to the local electrostatic potential of the RNA using Ionize [33]. The systems were carefully solvated in two phases with the TIP3P water model [62]: first, Solvate [35] was used to place the first solvation layers (8 Å), and, second, the VMD solvate plugin [36] to complete the water box with a minimum of 20 Å buffer region on each side. The resulting systems had sizes similar to 1,100,000 atoms.

MD simulations were performed using the latest version of NAMD 2.10 [37]. To guarantee correct local solvent density and ion solvation shell around the highly charged backbone and deep groove of the RNA molecules [32, 63, 64], all prepared systems were minimized and equilibrated in a step-wise fashion. Minimization was carried out using the conjugate gradient method in NAMD. Positional constraints were placed on all heavy-atoms for 2,000 steps. Afterwards, constraints were then released for the water molecules for 3,000 steps. Protein and nucleic acid side-chains, as well

as the ions were set free for the next 5,000 steps. Finally, all atoms were set free for the last 20,000 steps of minimization. Thermalization was conducted using a temperature jump protocol with step-wise positional restraints to allow waters and ions to diffuse slowly into and pack against the RNA structure. The initial temperature was set to 100K, and ions and heavy atoms in the RNA and protein were harmonically restrained for 25 ps. Then, the temperature was raised to 200K, and ions and the backbone atoms were harmonically restrained for 25 ps. In the next step, the backbone atoms were harmonically restrained at the temperature of 250K for another 50 ps. Force constants for all harmonic restraints were set to 1 kcal·mol^{-1} · Å^{-1}. Finally, the temperature was raised up to 300K and all atoms were freed for further equilibration. A total of 840 ns of MD simulation on the 16S intermediates were reported.

Production runs are conducted using NAMD 2.10 [37] under the NPT ensemble at 1 atm and 300 K. Periodic boundary conditions are applied, and a 1 fs - 2 fs - 4 fs multiple time-stepping approach was used. Long range interactions are calculated using PME with 10 Å switching/12 Å cutoffs. Each run uses approximately 40,000 node-hours on NCSA Blue Waters's XE6 nodes (2× AMD 6276 Interlagos).

2.3 Results

Using kinetic binding parameters, we developed an ODE model to describe assembly of the ribosomal small subunit [65]. From the ODE model, we analyzed several key intermediates to determine the functional role of several of the r-proteins in the assembly process.

		Bound r-proteins				
Index	States	Central domain	3′ domain domain	Number of atoms	Dimensions	Simulation time (ns)
1	200	-	-	1,046,000	$182 \times 202 \times 290$	140
2	201	-	S7	1,041,000	$181 \times 202 \times 289$	140
3	200:8	S8	-	1,027,000	$179 \times 201 \times 289$	140
4	200:15	S15	-	1,031,000	$179 \times 201 \times 290$	140
5	201:8, 19	S8	S7,S19	1,052,000	$180 \times 205 \times 290$	140
6	201:8, 9, 19	S8	S7,S9,S19	1,011,000	$176 \times 200 \times 290$	140

Table 2.1: Summary of MD simulations performed in this paper. All systems have the following 5′ domain r-proteins prebound: S4, S17, S20, and S16.

The minor pathway in the kinetic model has not been experimentally observed; however, the proteins bound to the *in vitro* states 100 and $200:8$, appearing before and after the bifurcation point, have been predicted using cryo-EM and P/C qMS [8]. Using MD simulations, we probed the ensemble of conformations 201, $200{:}8$, and $200{:}15$ near the bifurcation point at 200 (Table 2.1). All states contain the intact 16S rRNA and are prebound with S4, S17, S20, and S16 while states 201, $200{:}8$, and $200{:}15$ have in addition S7, S8, and S15 bound respectively. To observe the maximum fluctuations in the nucleic acid conformations, we prepared the MD simulations with a

neutralizing concentration of sodium ions with no magnesium ions present.

In our previous MD simulations and experiments [1,2,66] on motions of the 5′ domain under similar conditions, we saw that the dominant role of S4 in state 100 and 200 is to bring together helices h16 and h18, while (S17, S20, and S16) r-proteins tighten helices in their binding sites on the 5′ and central domain. Because the central domain is already partially formed in state 200, it is expected that the main role of S8 and S15 is to add rigidity to the central domain. S7 binds to the partially formed 3′ domain while S8, and S15 binds to regions in the central domain already formed (see Fig. 2.4 and 2.5 in the Supporting Material).

In the 3′ domain, all four simulations showed similar motions. These fluctuations are dominated by the partial unfolding of the 3′ domain. Helices in the lower four way junction (h29, h30, h41-h43) separate from helices in the upper three way junction (h34-h40) (Fig. 2.1a). Time traces of the center of masses for the different junctions in all four MD simulations show that the helices separate from 40 Å to over 60 Å after 140 ns (Fig. 2.1b). Simultaneously, the structural signature [57] h33 separates from h31 and h32 and becomes more solvent exposed. This is expected since h33 is connected to these junctions. Similar results are seen in simulations with the Thermus thermophilus (T. thermophilus) small subunit (Fig. 2.2 in the Supporting Material), suggesting that these motions are probably common to all bacterial organisms. The fact that states $200; 201; 200{:}8$ or $200{:}15$ all have similar motions suggests that there is no strong bias to binding either S7, S8, or S15 and that the next major assembly barrier, the opening of the 3′ domain,

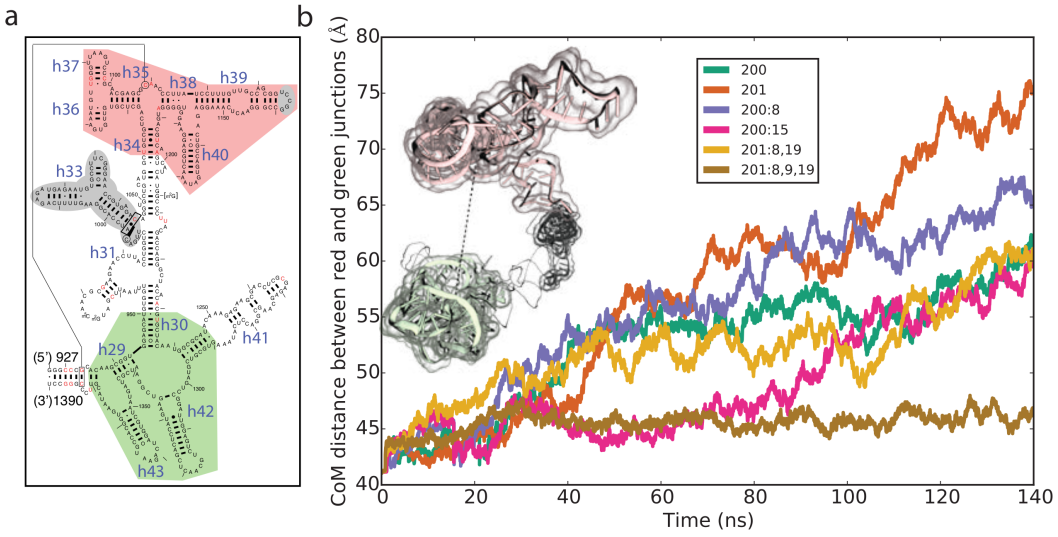

Figure 2.1: a) Secondary structure diagram of the 3′ domain with the center of masses defined. Center of masses are computed from the lower four-way junction helices h29, h30, h41-h43 (green) and the upper three-way junction helices h34-h40 (red). The exact residues are marked on the modified secondary structure diagram [56]. These centers are separated by the structural signature—marked in gray circles—h33 and numerous sequence signatures [57]. (b) Time traces of center of mass distances in the 3′ domain. The r-protein binding sites in the folded mall subunit, for each domain, are shown in Tables 2.3, 2.4 and 2.5 in the Supporting Material.

occurs further along in the assembly pathway.

Because the binding of S7 and S8 have a minimal global effect on the structure of ribosome assembly intermediate, we probed the effect of adding two 3′ domain binding r-proteins (S9 and S19). In the folded ribosomal small subunit, S9 binds to both the lower four way and upper three way junction while S19 binds to the structural signature h33 (Fig. 2.5). As the S19 binding site is more local than S9, we probed its binding first S19 (Fig. 2.5). Adding S19 to the simulations (moving from state $200{:}8$ to $201{:}8,19$) tightens the structural signature in h33 and keeps h33 packed against h31-h32. Like the

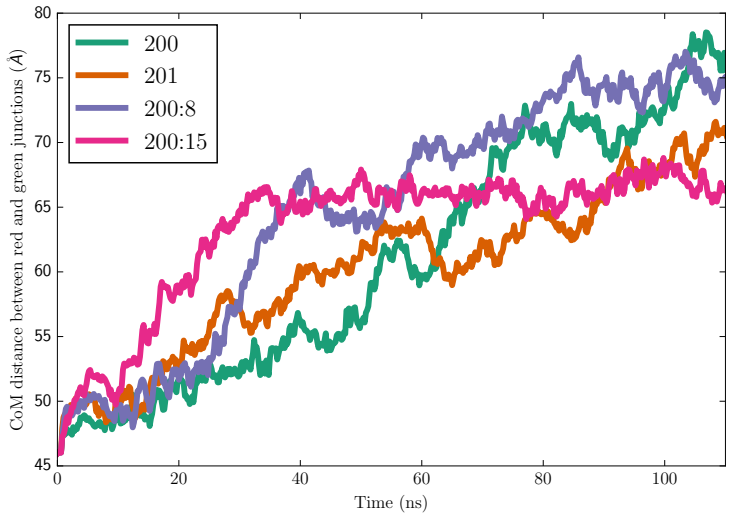

Figure 2.2: Separation of junctions in the 3′ domain of T. thermophilus. The figure is analogous to the one in the Main Document (Fig. 2.1). Starting conformation taken from the PDB 1HR0. The simulation protocol is identical to the one used for the *E. coli* simulation.

four previous simulations, state 201:8, 19 also shows similar unfolding of the 3′ domain (Fig. 2.1b). State 201:8, 9, 19, on the other hand, does not have the separation in the 3′ domain (Fig. 2.1b). Interestingly, all six MD simulations showed the 3′ domain rotating away from the five-way junction in the 5′ domain, suggesting that there is another folding barrier further along in the assembly pathway. This motion might only be arrested upon the addition of S5.

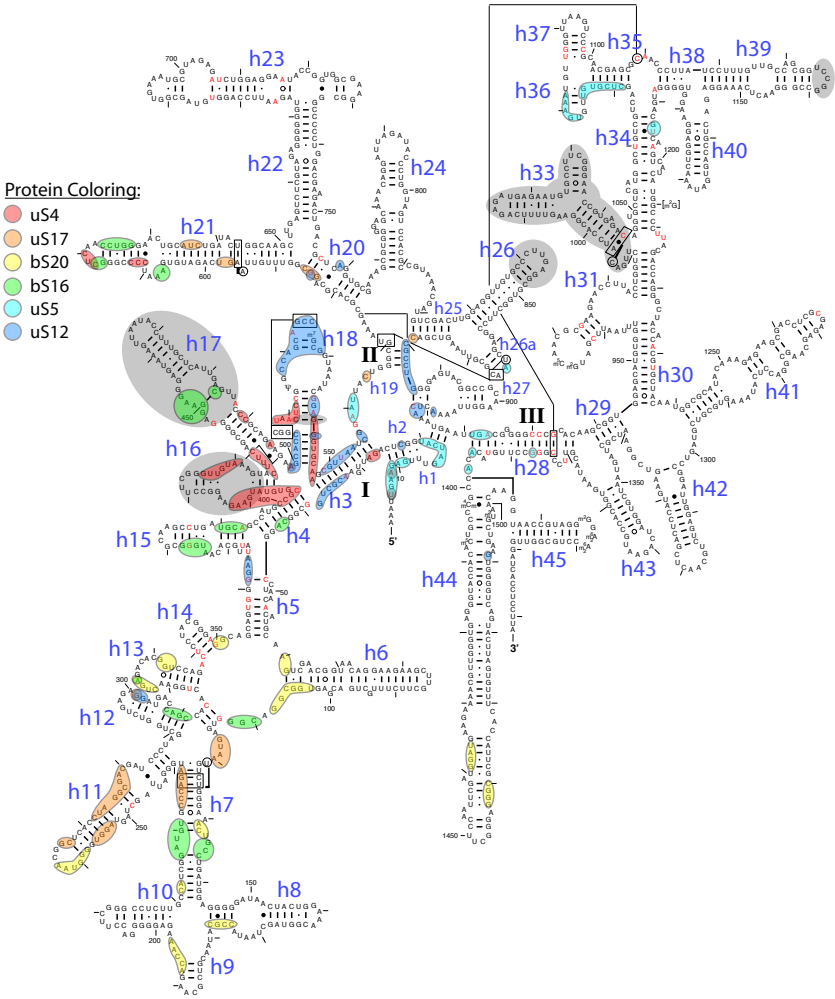

Figure 2.3: Secondary structure diagram of E. coli with central domain r-protein binding sites (in the folded 30S subunit) labeled. R-protein binding sites determined using a 5Å from the crystal structure 2I2P [28]. Red letters and gray shapes denote sequence and structural rRNA signatures respectively [57]. Map is based on 16S rRNA map from Cannone, et al. [44].

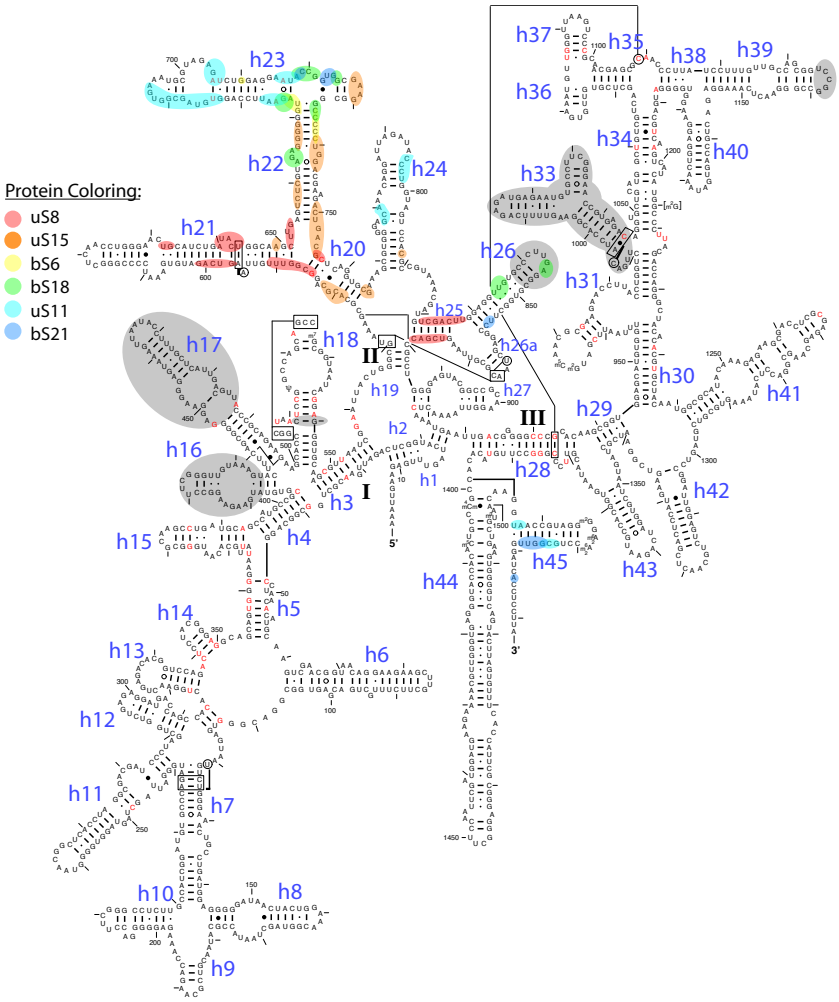

Figure 2.4: Secondary structure diagram of E. coli with central domain r-protein binding sites (in the folded 30S subunit) labeled. R-protein binding sites determined using a 5Å from the crystal structure 2I2P [28]. Red letters and gray shapes denote sequence and structural rRNA signatures respectively [57]. Map is based on 16S rRNA map from Cannone, et al. [44].

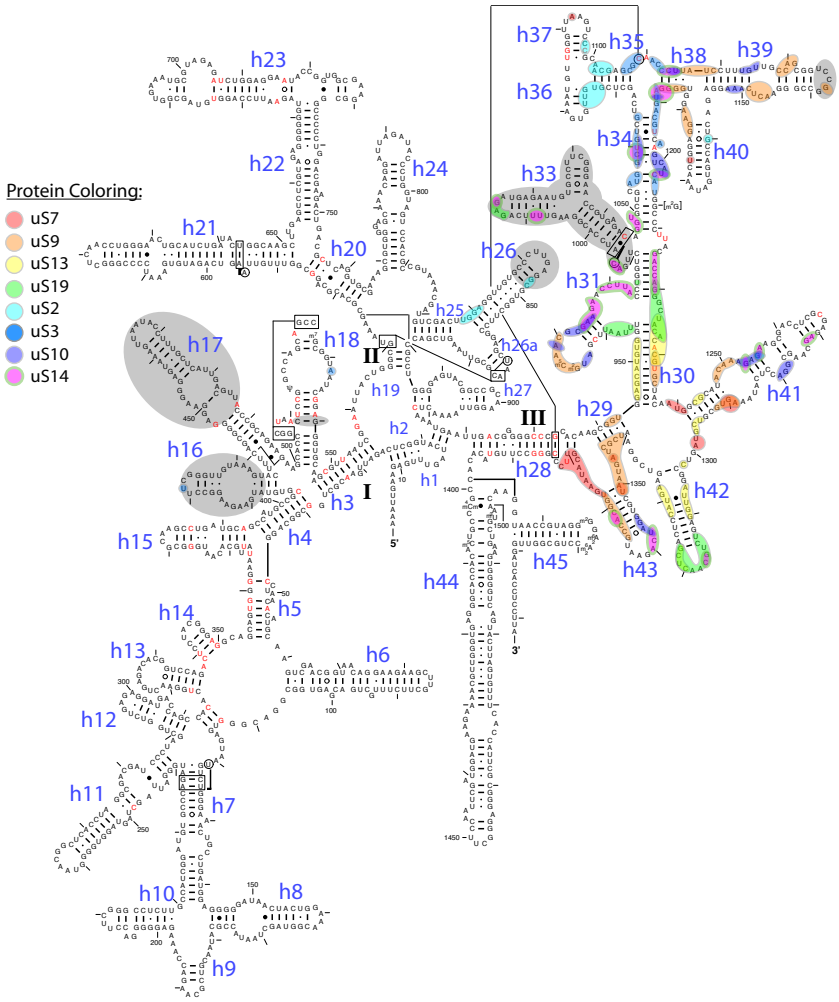

Figure 2.5: Secondary structure diagram of E. coli with 3′ domain r-protein binding sites (in the folded 30S subunit) labeled. R-protein binding sites determined using a 5Å from the crystal structure 2I2P [28]. Red letters and gray shapes denote sequence and structural rRNA signatures respectively [57]. Map is based on 16S rRNA map from Cannone, et al. [44].

Chapter 3

Elongation factor-Thermo unstable (EF-Tu)†

3.1 Abstract

Elongation factor Tu (EF-Tu) is a highly conserved GTPase responsible for supplying the aminoacylated tRNA to the ribosome. Upon binding to the ribosome, EF-Tu undergoes GTP hydrolysis which drives a major conformational change, triggering the release of tRNA to the ribosome. Using a combination of molecular simulation techniques, we studied the transition between the pre- and post-hydrolysis structures through two distinct pathways. We show that the lower free energy pathway involves separation of the GTP binding domain (Domain 1) from the OB folds (Domains 2 and 3), followed by Domain 1 rotation, and, eventually, locking EF-Tu conformation in the post-hydrolysis state. Using docking tools, we identified and characterized the EF-Tu conformations that release the tRNA. These calculations suggest that the EF-Tu can dissociate from tRNA before the domains dissociate and after Domain 1 rotates by 25°. We also examined the EF-Tu conformations in the context of the ribosome. Given the high sequence

†Work recently submitted and includes contributions from Zhaleh Ghaemi and Zan Luthey-Schulten. Material from the Supporting Information can be found here: https://uofi.box.com/v/JonathanLaiThesis2017

similarities with other translational GTPases, we predict a similar separation mechanism is followed.

3.2 Introduction

Cells assemble proteins by linking monomeric amino acids together in the ribosome. Crucial to supplying amino acids to the ribosome is a heterotrimeric GTPase, Elongation factor Tu (EF-Tu) in Bacteria—EF-1α in Eukaryota and Archaea. In the cytoplasm, EF-Tu forms a ternary complex with both GTP and aminoacylated tRNA (aa-tRNA). Association of the ternary complex to the ribosome and the anticodon of the aa-tRNA to the codon of the mRNA triggers GTP hydrolysis [67]. The release of the GTP γ-phosphate in EF-Tu induces a conformational change [68] which increases the dissociation constant of the ternary complex by three orders of magnitude—from the nanomolar [69] to micromolar range [70]. Because of the speed of translation [71], EF-Tu needs to rapidly switch from its pre- and to its post-hydrolysis conformations [72–75]. Once the aa-tRNA is released, a cognate or near-cognate aa-tRNA can rapidly move into the ribosome and participate in peptide bond formation; mismatched aa-tRNA are kinetically rejected from peptide bond formation [76,77].

X-ray crystallography and cryo-electron microscopy have captured structures of EF-Tu in complex with GTP or GDP ligands bound to tRNA and ribosomes [78–80]. Structures of Thermus aquaticus (T. aq.) and *E. coli* bacterial EF-Tu in both the pre- and post-hydrolysis states have been also

resolved using X-ray crystallography [72,73]. Comparison of the pre- and post-hydrolysis structures reveal three major structural differences (Fig 3.1): 1) using the OB folds (Domains 2 and 3) for alignment, the GTP binding domain (Domain 1) of the post-hydrolysis EF-Tu rotates by 90° with respect to the Domain 1 of the pre-hydrolysis state; 2) switch I, residues 41 to 63 (40 to 62 in *E. coli*) changes from an α-helix and adopts a β-sheet conformation; 3) switch II, residues 81 to 101 (80 to 100 in *E. coli*), partially unwinds and rotates approximately a quarter of a turn [73]. It is thought that the switch II favors the post-hydrolysis conformation, which can only occur when the hydrogen bond between the γ-phosphate of the GTP and H85 (84) of switch II is broken [81].

Experimental and computational methods have been extensively applied to study: 1) possible mechanisms of GTP hydrolysis in EF-Tu [82–85]; 2) accommodation of the tRNA into the A-site of the ribosome [86]; and 3) potential transient interactions between EF-Tu and other components involved with protein synthesis [31, 45, 46, 87, 88]. Although the pre- and post-hydrolysis states of EF-Tu are structurally known, the mechanism and timescale of its conformational change after hydrolysis and the EF-Tu structure that leads to the aa-tRNA release remains unknown [89].

Here, we used a combination of unbiased molecular dynamics (MD), coarse grained structure based models (SMOG), and enhanced sampling techniques—such as umbrella sampling (US) and steered MD (SMD)—to study pathways connecting the pre-hydrolysis (state **a**) to the post-hydrolysis (state **f**) states of EF-Tu. The pre-hydrolysis state refers to the protein con-

formation since the ligand used for all of the calculations is a GDP, namely, in the pre-hydrolysis structure PDB: 1B23) GTP has been mutated to GDP. We thoroughly sampled two distinct pathways: one directly connects state **a** to state **f** through the rotation of Domain 1 and the other involves a combination of Domain 1 rotation and the separation of Domain 1 from Domains 2 and 3. Based on our free energy calculations, we concluded that the latter pathway has a lower free energy barrier. This pathway together with its sampled structures were used in combination with docking techniques [90] to determine the EF-Tu conformation along the minimum free energy path (MFP) that releases the tRNA. Because the crucial residues for the EF-Tu conformational change are highly conserved among other translational GTPases, we suggest that our proposed mechanism is likely to be universal among all three domains of life.

Understanding the transition mechanism between the pre- and post-hydrolysis states of EF-Tu can help rationalize the improvement of antibiotics such as kirromycin and efrotomycin [91, 92] that are recently being administered to target the bacterial translational elongation factors in farm animals.

3.3 Methods

3.3.1 Molecular dynamics

Models of EF-Tu in the pre-hydrolysis (PDB: 1B23 [72]) and post-hydrolysis conformation (PDB: 1TUI [73]) were prepared using CHARMM36 [60, 61] for

protein/nucleic acid. Systems were ionized and solvated with a water buffer of at least 15Å using the protocol presented in Eargle et al. [51] (described in the Supporting Information). Simulations were performed in the NPT ensemble at 300K/1 atm using a 2 fs time step.

3.3.2 Preparing the transition pathways

To model the transition of EF-Tu from state **a** to state **f**, we constructed a heavy-atom SMOG model of EF-Tu [42]. The native potential was constructed using a cut-off based contact map of EF-Tu in state **f**. The default parameters were used for all other calculations. Simulations were performed with a 0.5 fs timestep and at a reduced temperature of 0.2 (25K) in the NVT ensemble. Frames were written every 5 fs. Twenty unique SMOG trajectories and velocity distributions were generated, and the RMSD distance between pairs of trajectories were calculated. The path that had the smallest pairwise distance to the other paths was chosen as our initial guess; the RMSD of the path approaches within 0 Å and 0.6 Å of the pre- and post-hydrolysis states, respectively.

Individual snapshots, separated by a RMSD of 1.7Å from each other, were extracted from the SMOG trajectory. Path-collective variables (S, Z) [93], as implemented in Plumed 2.2 [94], were used as a collective variable to describe the transition from state **a** to state **f**, using a lambda of 13. The atoms defining the path were as follows: Cα atoms of the Domain 1, heavy atoms of the switch I and amino acids at the interface of all three domains, and backbone atoms everywhere else. Constant velocity SMD [95] was

used to optimize the initial path through an iterative process in accordance to [96]. Briefly, over the course of 5 ns, the pre-hydrolysis state was pulled towards the post-hydrolysis conformation along the path prescribed by the SMOG trajectory. After each iteration, a new path was generated based on individual snapshots from the trajectory—resulting in the direct path between state **a** and state **f**. Care is taken to make sure that the snapshots are equally spaced in terms of RMSD and to capture the conformational changes described in the Introduction. Inspired by unbiased MD simulations (committor analysis) along the direct path, we generated the "separation path" by using SMD to pull Domain 1 from the other domains at S values of 4.8 and 18. Fig: 3.6 and 3.7 show the convergence of the optimization protocol for the separation and direct pathways, respectively.

To generate conformations that were orthogonal to the initial pathway (i.e. along the Z), we performed either well-tempered metadynamics [97] (low Z values) or extracted snapshots from the SMD (high Z values).

3.3.3 Free energy calculations

We used 2-dimensional umbrella sampling method along the S and Z coordinates. Umbrellas were placed on a regular grid along S (from 1 to 24 every 0.25) and Z (every 1 Å), totaling 306 umbrellas.

Each umbrella was restrained with a harmonic potential with a force constant of 25 kcal·mol^{-1}·S^{-1} and 2500 kcal·mol^{-1}·Å^{-1} (along Z), and equilibrated for 2 ns. Production run for each umbrella was at least 8 ns. The free energy surface was calculated with weighted histogram analysis method

(WHAM) [98] using the same grid spacing as the the umbrellas. To check the convergence of the calculation, the free energy was calculated from either the first 5 ns of the production run or from all 8 ns; the average difference between the two free energy surface was 0.45 kcal·mol^{-1}. Additionally, we checked the overlap of umbrellas and plotted the number of overlapping umbrellas in S and Z space (Fig: 3.16) as well as the density of states sampled for each S and Z bin (Fig: 3.17). To examine the obtained free energy we performed committor analysis by launching unbiased MD simulations from the two minima (Fig: 3.18) and from state **c** and state **d** (Fig: 3.14).

3.3.4 Estimating the binding of ET-Tu to tRNA

The EF-Tu bound to tRNA, with or without A76, was pulled along the MFP using constant-velocity SMD restraining Z$>$ 2 Å with force constants of 1000 kcal·mol^{-1}·S^{-1} and 1000 kcal·mol^{-1}·Å^{-1}. The protein and tRNA conformations were extracted from the SMD and interaction energies between the two were scored using Autodock Vina [99].

3.3.5 Correlations between EF-Tu and aa-tRNA

Correlations were generated using the protocol based on Van Wart, et. al [100] (see Supporting Information for details).

3.3.6 Generation of sequences and conservation

GTP-binding proteins (GO:0005525) are extracted from the UniProt database and clustered based on their titles, discarding clusters with fewer than 20 entries. These clusters are as follows: EF-Tu/EF-1A: 925, EF-G: 30, RF2: 53, Eef2: 51, LepA: 3545, Guf1: 93, RRF3: 179, IF2: 1044, and Era: 2912. Curated sequences of Ras (mouse and human) and selenocysteine-specific elongation factor (SelB *E. coli*), from Swiss-Prot, are added to the sample dataset *ex post facto*. Clusters are aligned to the reference EF-Tu sequence in PDB: 1B23 using MAFFT-L-INS-i [101] (ver. 7) with default parameters. Mutual information was calculated using Weblogo [102].

3.4 Results and Discussion

3.4.1 Pre- to post-hydrolysis conformational change of EF-Tu involves separation of its domains

To generate the initial path connecting the pre- to post-hydrolysis states, state **a** and state **f** respectively, twenty SMOG trajectories were launched—starting from the pre-hydrolysis crystal structure (PDB: 1B23). An exemplar path from the trajectories was selected and iteratively optimized [96] (see Methods for details) (Fig: 3.6, 3.7). We used path-collective variables [93] to describe the conformational change both along (S, a dimensionless quantity denoting conformational change progress) and orthogonal (Z, distance from a path) of the optimized pathway. We sampled two distinct pathways: the

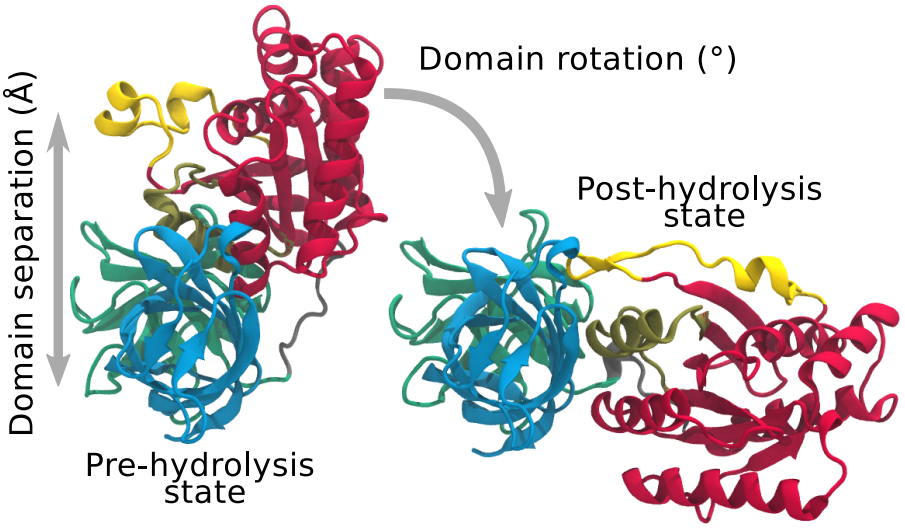

Figure 3.1: EF-Tu transition from the pre- (left) to post-hydrolysis (right) conformations. The conformational change largely involves a rotation of GTP binding domain Domain 1 (red) about OB folds Domains 2 and 3 (green and blue, respectively). Regions such as the switch I (yellow) and switch II (tan) also change secondary structure. The domain angle is defined as a dihedral angle between the center of mass of the Domain 1, Domain 2, linker (residues 218 to 220), and Domain 3. Only backbone atoms (C, Cα, N, O) were used to define the center of masses. Domain separation is defined as the distance between the center of mass of Domain 1 and 3.

first one directly connects state **a** and **f** based on Domain 1 rotation (direct pathway) (Fig. 3.1); the second one includes the separation of Domain 1 from Domains 2 and 3 in addition to the domain rotation (separation pathway).

The direct pathway can be divided into three segments: initial conformational change (**a** $\rightarrow$ **b**), large scale rotation (**b** $\rightarrow$ **e**), and final conformational change (**e** $\rightarrow$ **f**) (Figure. 3.2A, red dashed line). The separation pathway, on the other hand, can be divided into five parts: (**a** $\rightarrow$**b**), separation of the domains (**b** $\rightarrow$**c**), free diffusion of the Domain 1 (**c** $\rightarrow$**d**), rejoining of domains (**d** $\rightarrow$**e**), and (**e** $\rightarrow$**f**) (Fig. 3.2A, gray dashed line). Both pathways

approached within 1.08 Å backbone RMSD of the post-hydrolysis crystal structures (Fig. 3.8). The amino acids that deviated the most from the crystal structure are found either in the flexible region of the switch I or in loops connecting the secondary structure elements in Domain 1 and 2 (Fig. 3.8). The free energy surface is calculated using 2D-umbrella sampling along both the S and Z path-collective variables [93] (Fig. 3.9, direct pathway; 3.10, separation pathway), and remapped along the more intuitive collective variables of domain rotation and separation—shown in Fig. 3.2.

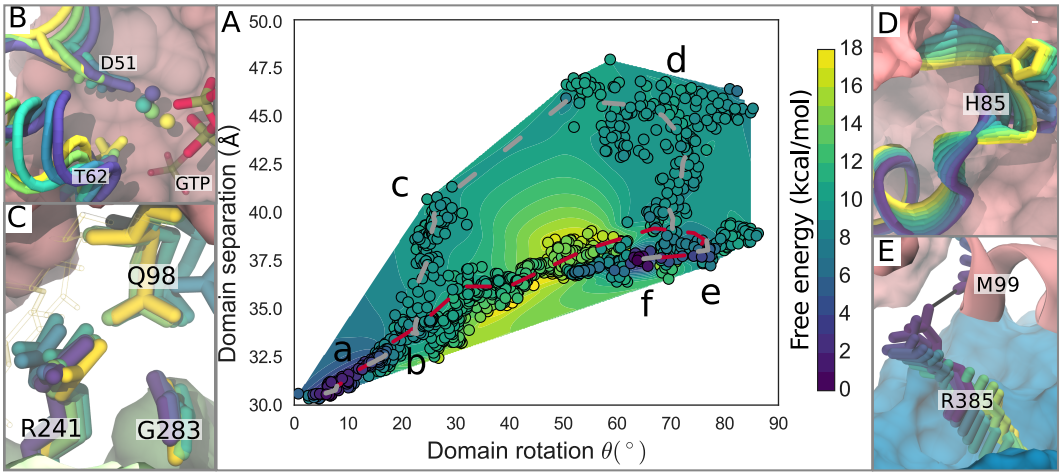

Figure 3.2: A) Free energy surface along domain rotation and domain separation. Free energies are calculated along S and Z path-collective variables and are remapped onto the space of domain rotation and the domain separation. Data from umbrella sampling calculations are shown in circles; the surface between states **c** and **d** is interpolated. The conformational changes from B) state **A** to **B**, C) **B** to **C**, D) **C** to **D**, and E) **D** to **E** are shown. Multiple snapshot from the MFP are colored from yellow to purple, towards progression to the post-hydrolysis state. Transparent molecules indicate the location of residues in the PDB: 1B23.

The free energy surface suggests that the difference between state **a** and **f** is approximately -1 ± 0.45 kcal·mol^{-1} which agrees with the experimentally

measured $\Delta\Delta G \approx$ -2.5 kcal·mol^{-1} between EF-Tu bound to GTP and GDP, respectively [103, 104].

The highest pre- to post-hydrolysis transition barriers are around 10 kcal·mol^{-1} and 15 kcal·mol^{-1} for the separation and direct paths, respectively. Because of the lower barrier of the separation pathway, it is the system preferred transition pathway, along which fewer non-native contacts between Domain 1 and Domain 3 with respect to the direct path are formed. Below, we describe the details of the conformational changes along the separation pathway. Amino acid numbering is based on *T. aq.* structure (PDB: 1B23) following by *E. coli* residue numbers in parenthesis.

State a to b

In the pre-hydrolysis crystal structure (PDB: 1B23), a key Mg^{2+} is coordinated by both the β and γ phosphates of the GTP ligand as well as T62 (61) of the switch I region. Upon GTP hydrolysis, T62 releases the Mg^{2+}, causing residues A53—I61 of switch I to partially open up and become more solvent exposed. Concomitantly, the Mg^{2+} moves in between the α and β phosphates of the GDP ligand (Fig: 3.2B), where other nearby oxygen atoms—namely the side chains of T25 (25) and D51 (50), and water molecules—can coordinate with Mg^{2+} and complete its solvation shell. Direct coordination of D51 to the Mg^{2+} pulls N41—I50 towards the GDP binding site, allowing A53—I61 to pivot out further into the solvent.

From the free energy calculations (Fig: 3.2A), the conformational change from **a** to **b** leads to a 20° rotation of Domain 1 and a domain separation

of about 3Å. The free energy difference for the conformational change is 10 kcal·mol^{-1}—arising from breakage of more than 80% of the native hydrogen bonds between Domains (Fig: 3.11). In addition, R241 (230) switches hydrogen bonding partner from D100 (99) to Q98 (97). To examine this part of the free energy surface, we launched short unbiased MD simulations from **b**: all simulations rush towards the pre-hydrolysis state, confirming that the initial movement of EF-Tu is energetically expensive. We also performed a long-time scale simulation of EF-Tu with the more flexible CHARMM22* force field, starting from **a**, to check if **a** is truly a minimum on the free energy surface; after 1μs, the system remained mainly in state **a** (Fig. 3.12).

State b to c

As the GTP binding Domain 1 separates from Domain 2, the remaining inter-domain hydrogen bonds continue to break. The final breaking of the hydrogen bonds occur between the side chain of Q98 (97) and R241 (230), and the backbone of G283 (271) (Fig: 3.2C). Calculations with Autodock Vina [99] show that once the the center of the Domain 1 moves beyond 38 Å from the center of Domain 3 (Z >4Å), the interaction energy between the domains trends towards zero (Fig. 3.13).

State c to d

In this step, the rotation of Domain 1 towards the post-hydrolysis conformation completes. Because the domains are not interacting at state **c**, the conformational change resembles an iso-energetic diffusion (Fig: 3.2A,3.10).

Unbiased MD simulations launched from either **c** or **d** states freely moved towards the other endpoint (Fig. 3.14), confirming that the transition from **c** to **d** is unobstructed. Interestingly, this is where we predict the helix containing H85 (84) of switch II partially unwinds, adjusting to the post-hydrolysis crystal structure (Fig: 3.2D).

State d to e

As Domain 1 rejoins Domain 3, switch I region adopts the extended β sheet conformation seen in the post-hydrolysis crystal structure, allowing the positively charged residues (e.g. R59 (58)) in the switch I to bind to a negative patch on Domain 3 (Fig: 3.2E). Simultaneously, R385 (373) of Domain 3 docks in-between the switch II and adjacent α-helix (I120—V128) of Domain 1 and hydrogen bonds to the backbone of the switch II (Fig: 3.2E). On the separation pathway, the docking of R385 can occur quite easily because the domain separation minimizes any potential non-native contacts that would interfere with R385 during the domain rotation. On the direct pathway, however, R385 tends to sterically clash with I120—V128 as the domains rotate; the entanglement of R385 contribute to the large transition barrier seen on the direct pathway.

State e to f

As EF-Tu converges near the post-hydrolysis state, switch II can lock into its final position. The switch II region contains two highly conserved residues: P83 (82) and Y88 (87) (Fig. 3.3A). From state **e** to **f**, Y88 is initially

solvent exposed (purple) (Fig. 3.3A). As Domain 1 adjusts its conformation, Y88 slips in between residues Y70 to K90 and binds to a hydrophobic pocket which acts as a lock for the post-hydrolysis conformation. This pocket can only form if P83 moves away from Domain 1 and occupies the same location of T62 (pink) as found in the GTP bound state. Unbiased MD simulations launched from the pre- and post-hydrolysis crystal structures (PDB: 1B23 and PDB: 1TUI, respectively) show that the Y88 exists either a solvent exposed or solvent shielded state—with minimal overlap between the states.

3.4.2 EF-Tu releases from the $5'$ end of the aa-tRNA prior to domain separation

Our calculations predicted that the conformational change of EF-Tu involves the rotation of Domain 1 and separation of Domain 1 from Domains 2 and 3, and these global changes might be involved in the release of tRNA to the small subunit of the ribosome. To answer the question of when EF-Tu releases the tRNA, we used SMD to pull an equilibrated EF-Tu and aa-tRNA complex along the minimum free energy path (MFP) and estimated the protein-nucleic acid binding energy with Autodock Vina [99]. The calculations suggest that the 5′ end of the aa-tRNA acceptor stem is the first component to detach (Fig. 3.4A) leading to a loss of approximately 3.5 kcal·mol^{-1} of an interaction energy (Fig. 3.4B). As the domain rotation increases to 25°, the EF-Tu aa-tRNA interaction energy tends to zero (Fig. 3.4B).

As the protein changes conformation from state **a** to **b**, the correlations between EF-Tu to the acceptor stem and the T-stem decrease, following by

the increase of the correlation between the EF-Tu and T-stem for the state **b** to **c** transition (Fig. 3.4C). The correlation between Domain 2 and A76 mainly increases during both transitions. These results support the idea that 5′ end of aa-tRNA acceptor stem detaches from EF-Tu first followed by the T-stem and then A76 nucleotide.

In all of the SMD simulations, A76 remains hydrogen bonds to E271 (259) while the backbone of the amino acid charged on the A76 can transiently bind to N285 (273). Moreover, the rotation of EF-Tu moves most of the interacting residues, as identified in Eargle, et al [31], away from the aa-tRNA (Fig: 3.15).

After state **b**, no energetic contributions are seen from the amino acid side chain charged on the A76. Instead, intramolecular hydrogen bonding forces the ester-linked amino-acid to rotate towards C75, decreasing the overall solvent accessibility surface area for the entire aminoacylated-A76 residue. Because the acyl-linkage of the aa-tRNA is vulnerable to hydrolytic attack, Domain 2 probably protects the aminoacylated bond during the release of aa-tRNA into the ribosome. We also repeated the same SMD simulations without A76 to accelerate the aa-tRNA dissociation event. These simulations show that, the aa-tRNA easily dissociates from EF-Tu, suggesting that the rate-limiting step for aa-tRNA release is the interaction between Domain 2 and A76. Similar analysis on estimating the interaction energy between the EF-Tu and ribosome shows no appreciable energetic contribution from the ribosome (data not shown). In fact, snapshots of EF-Tu from the MFP docked into the ribosome show that Domain 1 has sufficient space to move away from the ribosomal large subunit during the conforma-

tional change (Fig: 3.15).

3.4.3 Universality of mechanism among all translational GTPases

Given the sequence and structural conservation of Domain 1 and 2 amongst translational GTPases in all domains of life [106, 107], the amino acids involved in the EF-Tu conformational change can also be involved in the mechanistic function of other translational GTPases. To test this possibility, we extracted and aligned more than 8835 translational GTPases sequences from all domains of life (Fig. 3.5). As expected, key residues in either switch I or switch II, such as D51 and T62 (state **a** to **b** transition), H85 (state **d** to **e**), P83 and Y88 (locking the post-hydrolysis state), are present in all GTPases (Fig: 3.5). Interestingly, residues Q98, D100 (or a negatively charged amino acid), and R241—all predicted to be involved in the **a** to **b** transition—and R385—**d** to **e** transition—are conserved amongst other GTPases (Fig: 3.5).

3.5 Conclusion

In conclusion, we have sampled two pathways that connect the pre- and post-hydrolysis states of EF-Tu. The pathway with lower free energy involves separation of Domain 1 from Domains 2 and 3 that allows Domain 1 to rotate freely to the final post-hydrolysis conformation. Following this pathway, we determined the EF-Tu conformation that dissociates from the tRNA and confirmed the feasibility of our suggested pathway by docking the EF-Tu

bound to the tRNA to the ribosome. Finally, based on the sequence similarity of EF-Tu to other translational GTPases, we argue that the conformational change mechanism of EF-Tu can be a universal process.

3.6 Supporting Information

3.6.1 Molecular dynamics

Models of EF-Tu bound to GTP in the pre-hydrolysis (PDB: 1B23 [72]) and with GDP in the post-hydrolysis conformation (PDB: 1TUI [73]) are prepared using the CHARMM 36 [60, 61] force field for protein and nucleic acid. Systems were ionized and solvated following Eargle et al. [51]. Briefly, the crystal structures were neutralized by placing K^+ions at electrostatic minima near the protein (using the Ionize [33]). Afterwards, the system is solvated using Solvate1.0 [35] and VMD 1.9.3 [36] to ensure that all dications have a full solvation shell and to maximize the protein-solvent interactions. The models have a minimum water buffer of at least 15 Å in all directions. Additional $K^+$and Cl^-ions were added to bring the total salt concentration to 0.150 M. Models contain approximately 95,000 atoms and have an approximate box size of $100\times100\times100$Å.

Simulations were performed in the NPT ensemble by using either a Langevin (with a coupling constant of 1ps) or Nose-Hoover [108, 109] thermostat set at 300K and pressure at 1 atm Berendsen barostat. Long-range electrostatics are calculated using the Particle-Mesh Ewald algorithm [110] with a cutoff of 12Å. SETTLE [111]/LINCS [112] is used and the timestep is

set to 2 fs. All simulations are performed using either NAMD 2.10 [37] and Gromacs 5.1 [113].

3.6.2 Correlation calculations

Generalized correlations [114] are calculated from the SMD trajectory. Frames are saved every 2 fs. Nodes were defined on the center of mass for amino acids and nucleic acids in accordance to the protocol in Van Wart, et. al [100]. Briefly, correlations are calculated over 1 ns trajectory segments and multipled by a 4.5$\mathring{A}$ contact map of the pre-hydrolysis state to get only the correlations between EF-Tu and the aa-tRNA. Finally, correlations connecting the EF-Tu and different parts of the aa-tRNA are averaged together.

3.6.3 Figures generation

All figures are generated using either Matplotlib [115], Seaborn [116], Weblogo [102], and VMD [36] or custom code written in Python 2.7/3.4 or Tcl/tk 8.5.

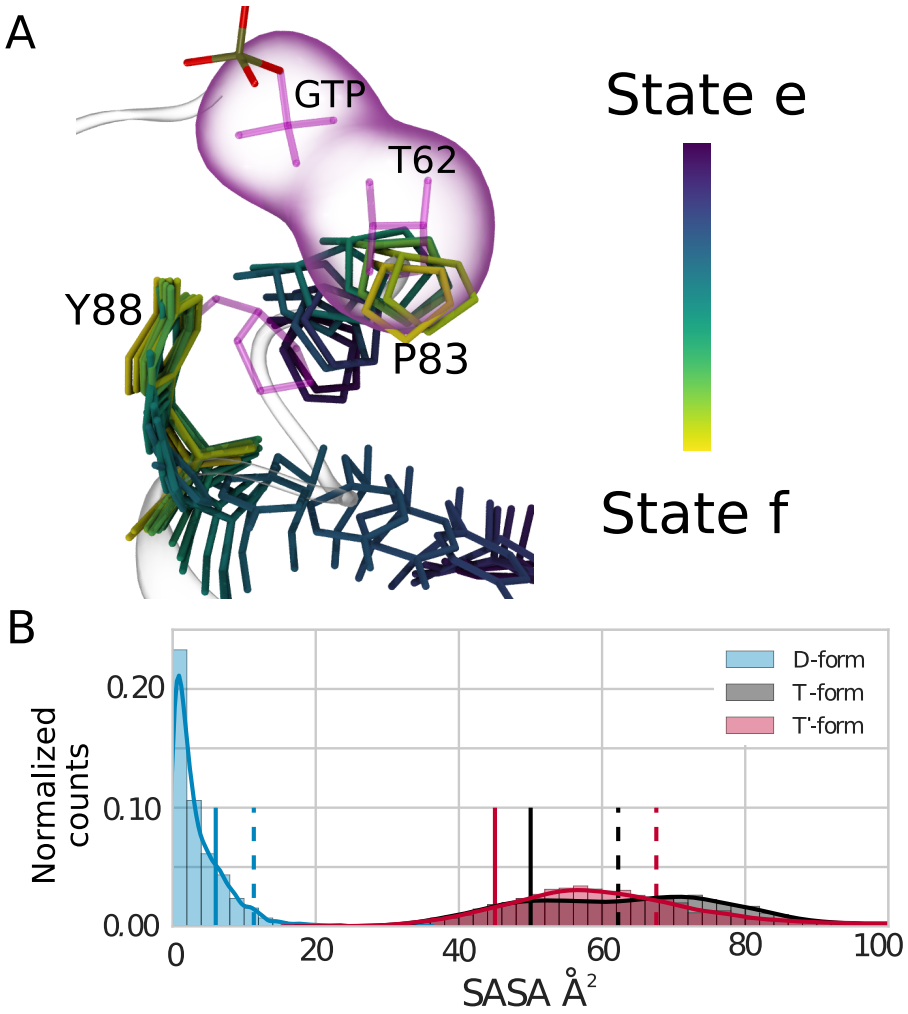

Figure 3.3: A) state **e** to **f**. Residues P83 and Y88 rearrange as the protein switches from the pre- (dark purple) to post-hydrolysis conformation (yellow). P83 and Y88 move into the void formed by the departure of the γ-phosphate and a formerly coordinated T62 in the GTP state (pink) respectively. B) Solvent accessibility surface area (SASA) of Y88 decreases drastically from the pre-hydrolysis conformation (with either GTP (T-form) or GDP (T′-form) bound) to the post-hydrolysis conformation (D-form). Solid and dashed lines indicate the SASA of Y88 at the beginning and end of the MD simulation, respectively.

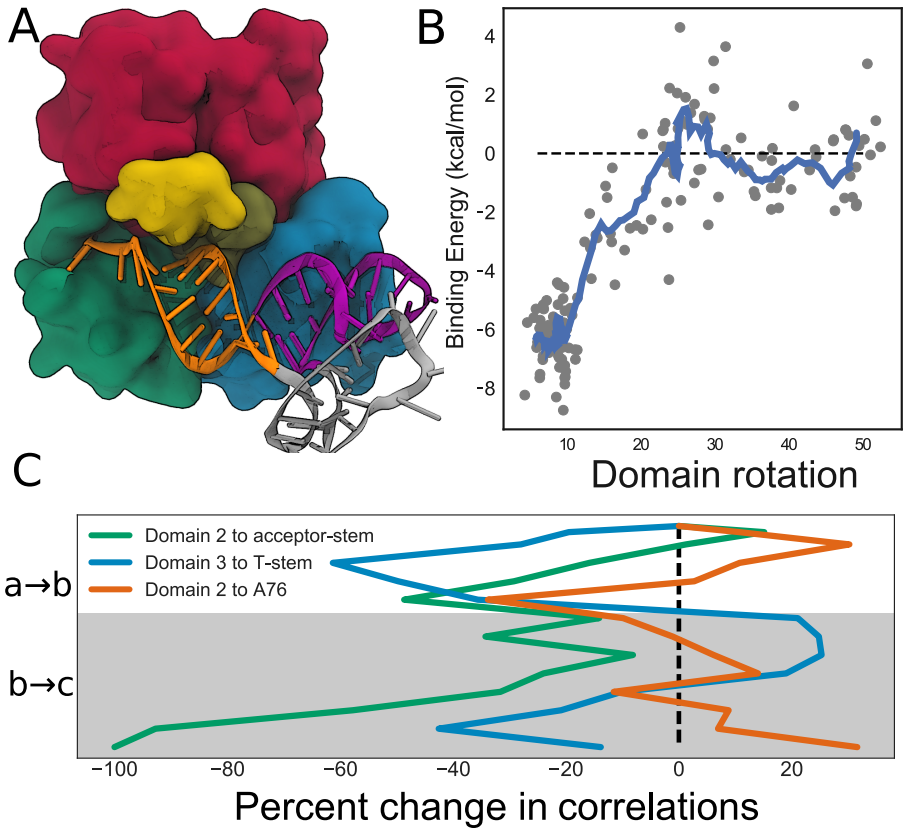

Figure 3.4: A) Initial structure of EF-Tu bound to tRNA; tRNA is colored as follows: acceptor-stem (orange), T-stem (purple), and D-loop (gray). B) Interaction energy between EF-Tu · aa-tRNA as calculated with Autodock Vina [99] vs. Domain 1 rotation angle. The interaction energies baseline was the value obtained for the post-hydrolysis crystal structure (gray line). Below 10° of Domain 1 rotation, EF-Tu binds to the tRNA around -6 kcal·mol^{-1}. As the Domain 1 continues to rotate, the binding between EF-Tu and aa-tRNA weakens and partially dissociates away. Experimentally, the binding energy of EF-Tu to the tRNA is -10.4 kcal·mol^{-1} [105], suggesting that Autodock underestimates the binding affinity by 4 kcal·mol^{-1}. C) Correlation values between the EF-Tu and aa-tRNA from the SMD simulations. Results shown for **a** to **b** (white) and state **b** to **c** transitions (gray).

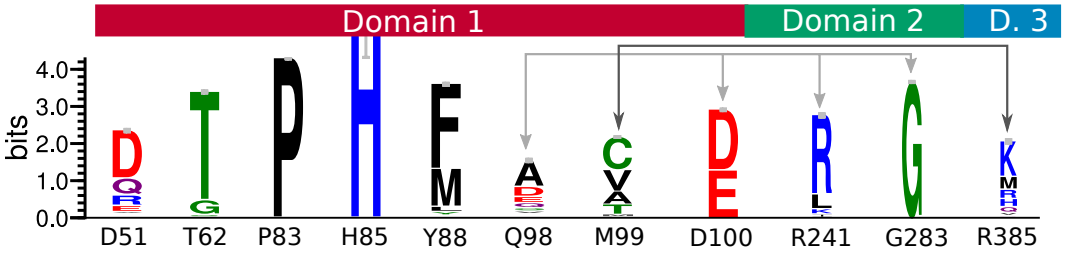

Figure 3.5: Conservation of select residues from the minimum free energy pathway amongst 8835 translational GTPases. Label indicate the residue in EF-Tu and the mutual information of each amino acid is plotted. Amino acids are colored by type: positively charged (blue), negatively charged (red), hydrophobic (black), hydrophilic (green), and other (purple). Arrows indicate interaction between residues.

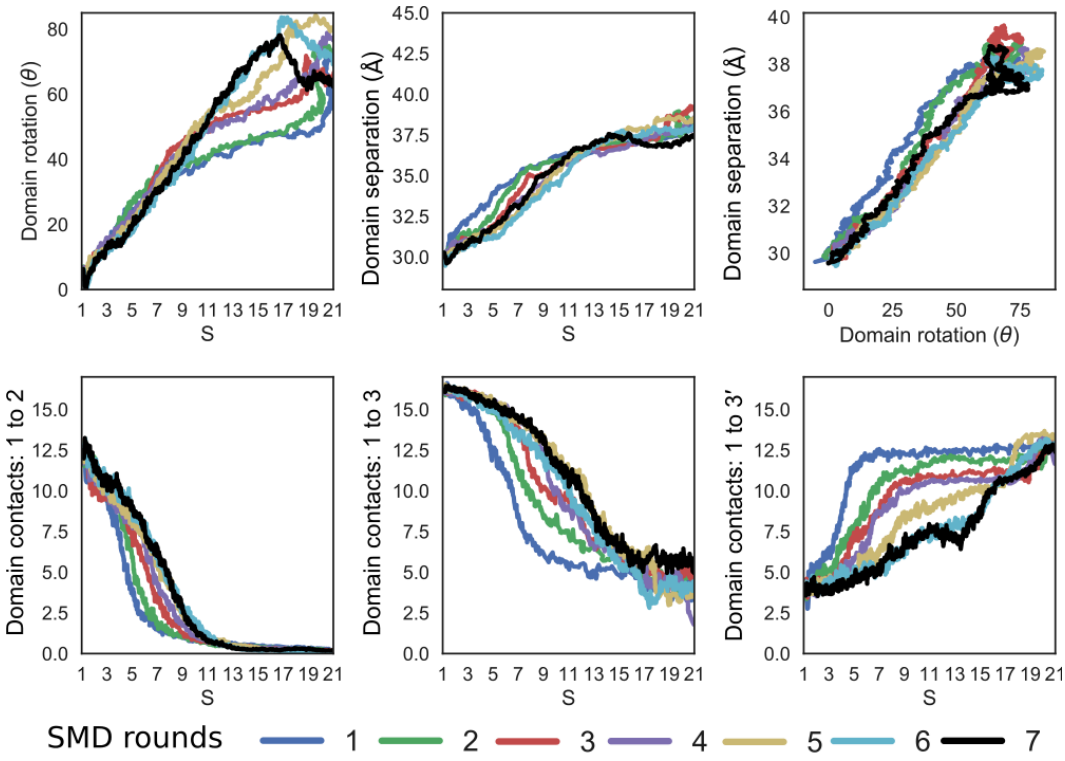

Figure 3.6: Convergence of the direct path using SMD. Pathways are iteratively optimized using the protocol described in the Methods section. Six different metrics were monitored: 1st) domain rotation, 2nd) domain separation, 3rd) domain rotation and separation, 4th) contacts between Domain 1 to Domain 2, 5th) contacts between Domain 1 to Domain 3 in the pre-hydrolysis state, 6th) contacts between Domain 1 to Domain 3 in the post-hydrolysis state.

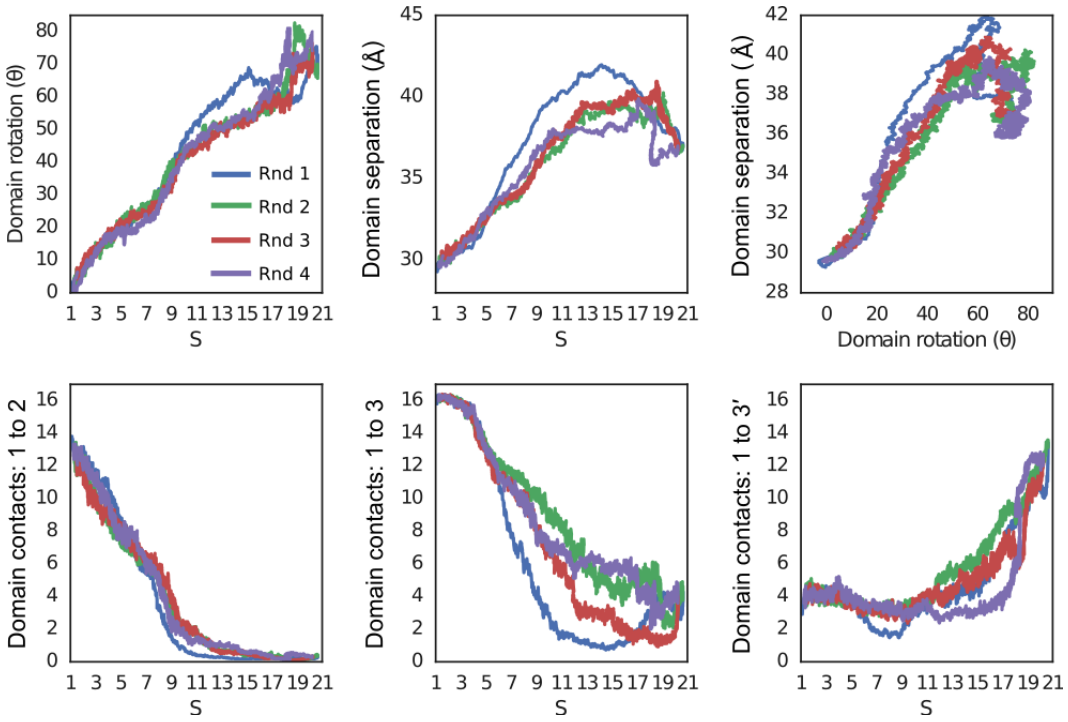

Figure 3.7: Convergence of the separation path using SMD.

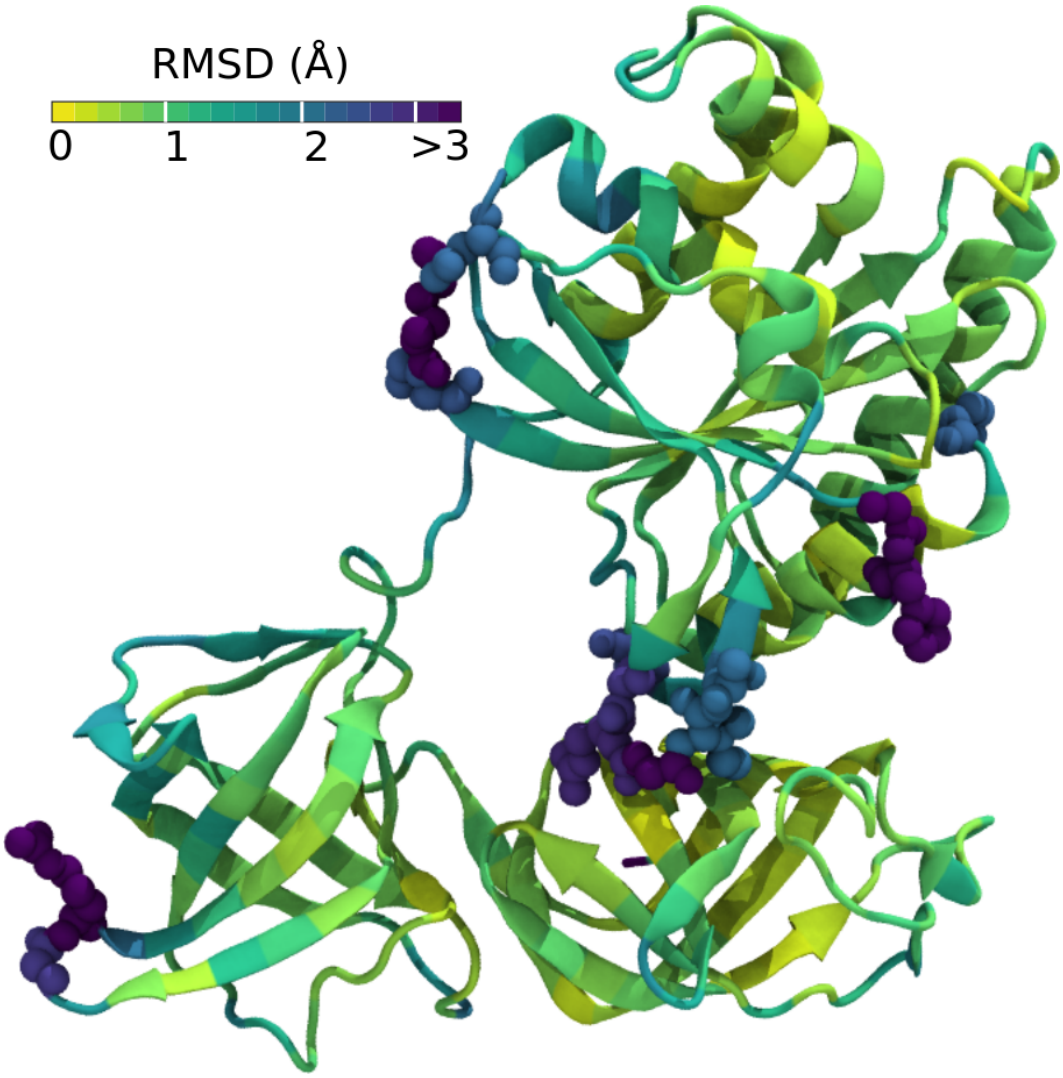

Figure 3.8: Structural comparison between simulated and crystalized post-hydrolysis EF-Tu. Amino acids are colored by backbone RMSD from 0 (yellow) to $>$ 3Å (purple). Amino acids with backbone RMSDs greater than 2Å are also shown in VDW representation. The average backbone RMSD is 1.08Å.

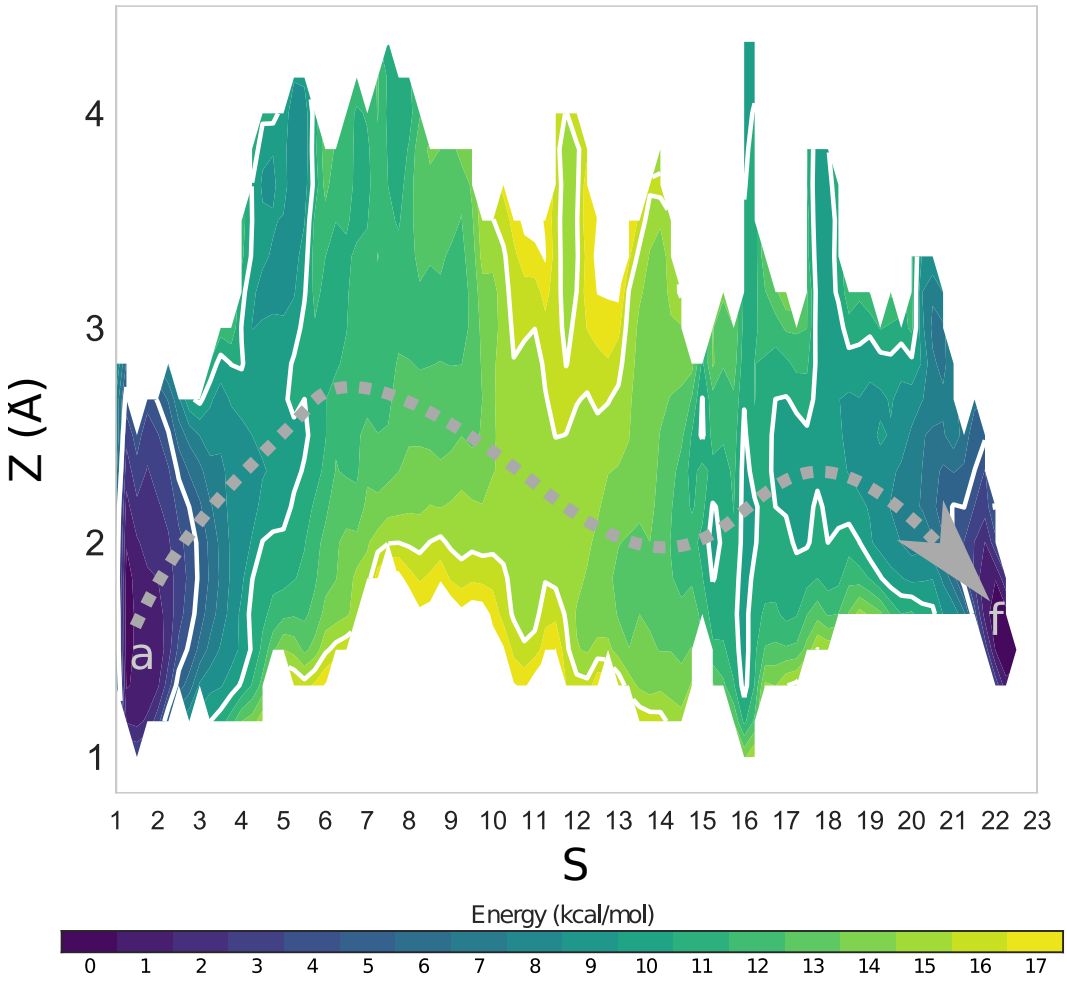

Figure 3.9: Free energy surface for the EF-Tu conformational change plotted in S and Z space for the direct pathway. Gray line shows the direct MFP.

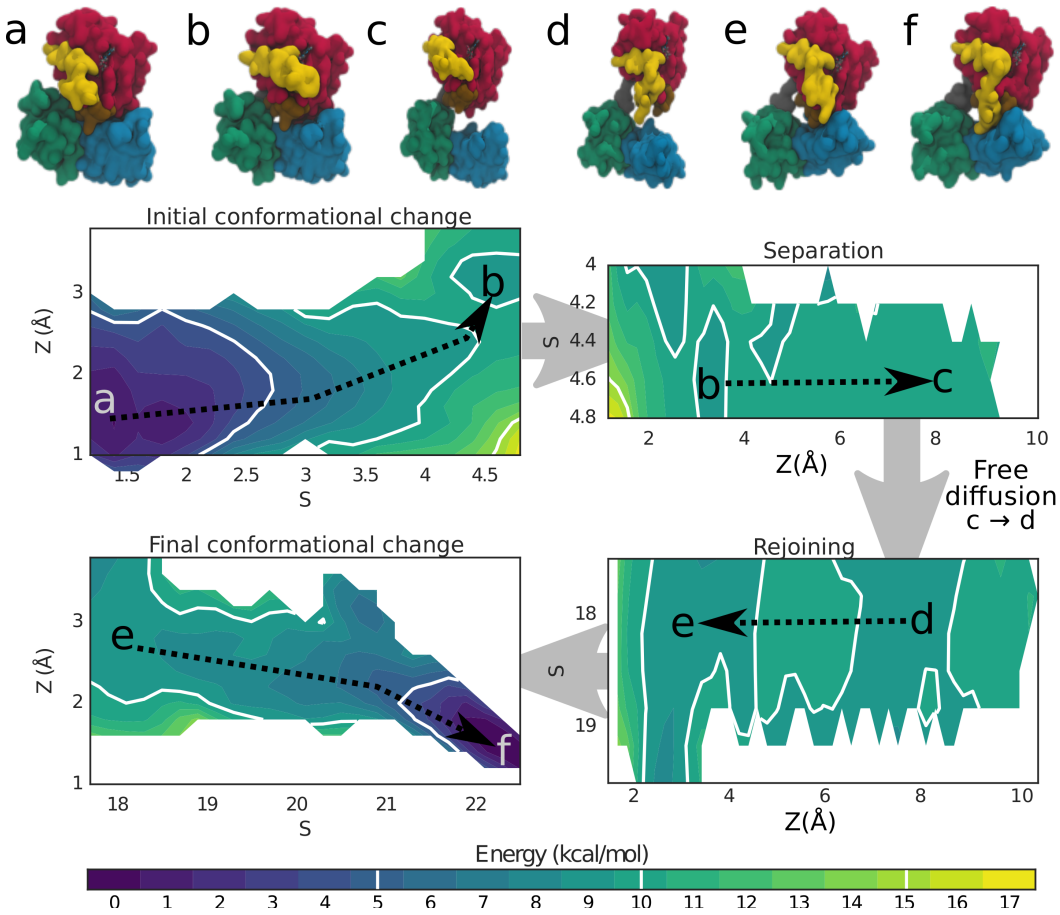

Figure 3.10: Free energy surface for the EF-Tu conformational change plotted in S and Z space for the separation pathway. The process is decomposed into four steps: initial conformational change, separation, rejoining, locking to the post-hydrolysis state. Snapshots for states a-f are shown above; Domain 1 (red), switch I (yellow), switch II (tan), OB-folds (green/blue).

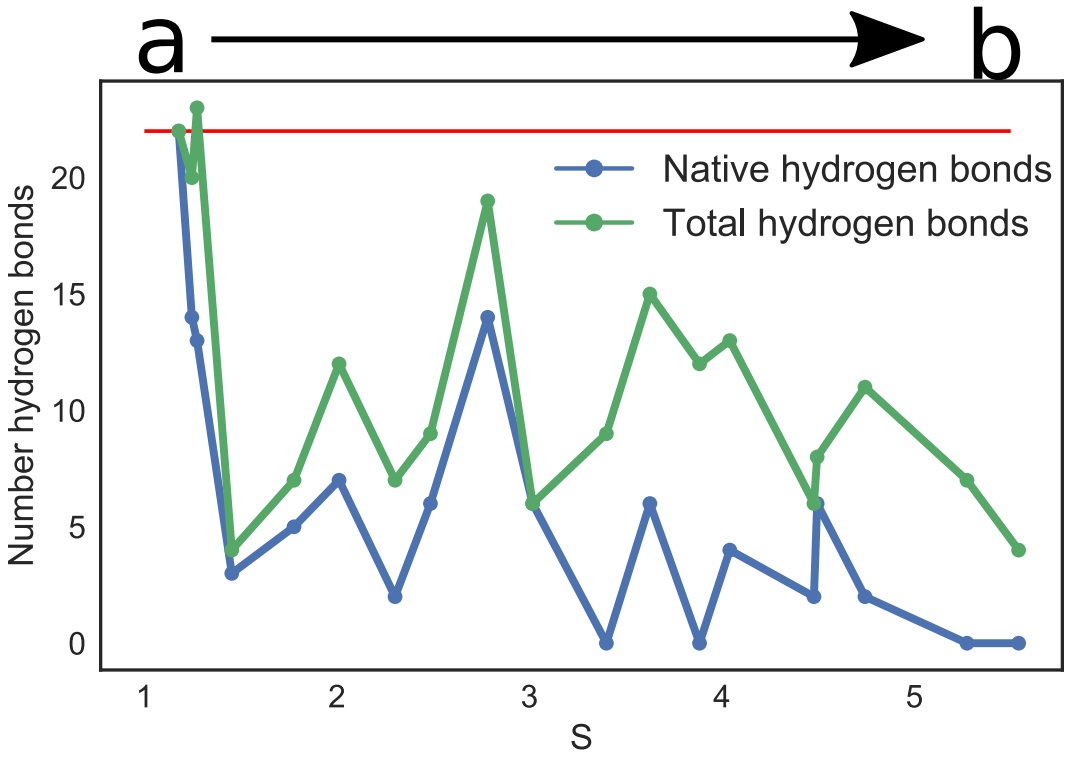

Figure 3.11: Interdomain hydrogen bonding of EF-Tu as a function of S. Hydrogen bonds are defined using a 3.5Å distance cutoff along the MFP. There are 22 hydrogen bonds in the equilibrated pre-hydrolysis conformation (red line).

.

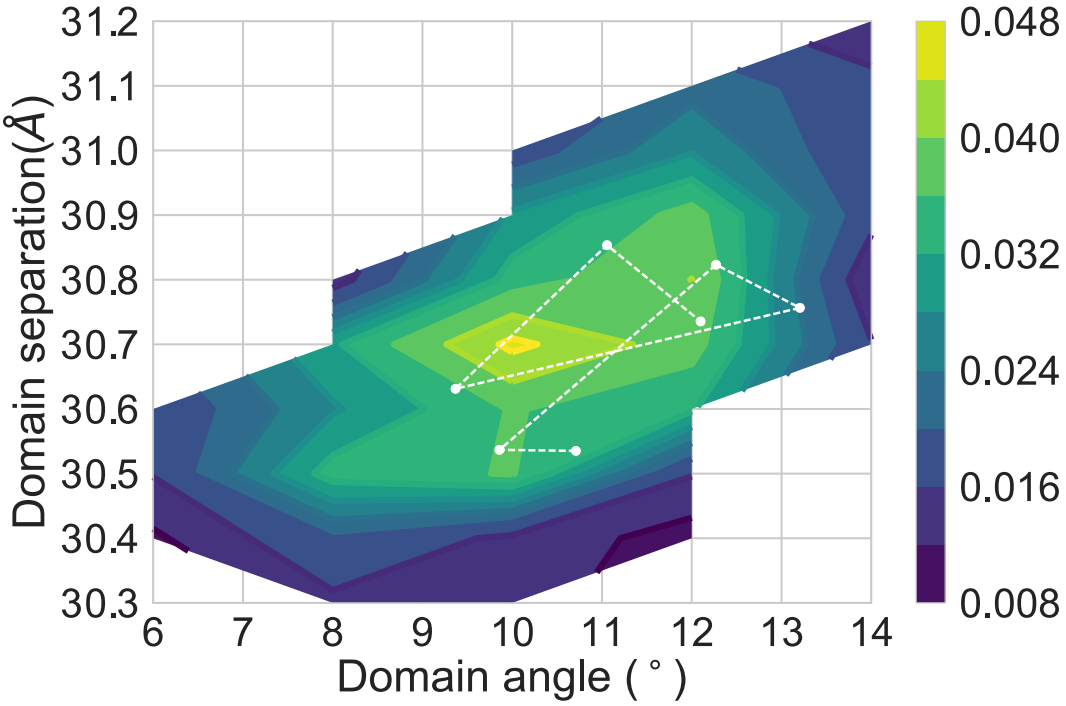

Figure 3.12: 1 μs simulation of the post-hydrolysis EF-Tu starting from 1B23 using the Charmm22* force field. A 2D-histogram in domain angle and separation space shows that the protein spends the majority of its time in state **a**, confirming that the transition to state **b** is energetically costly.

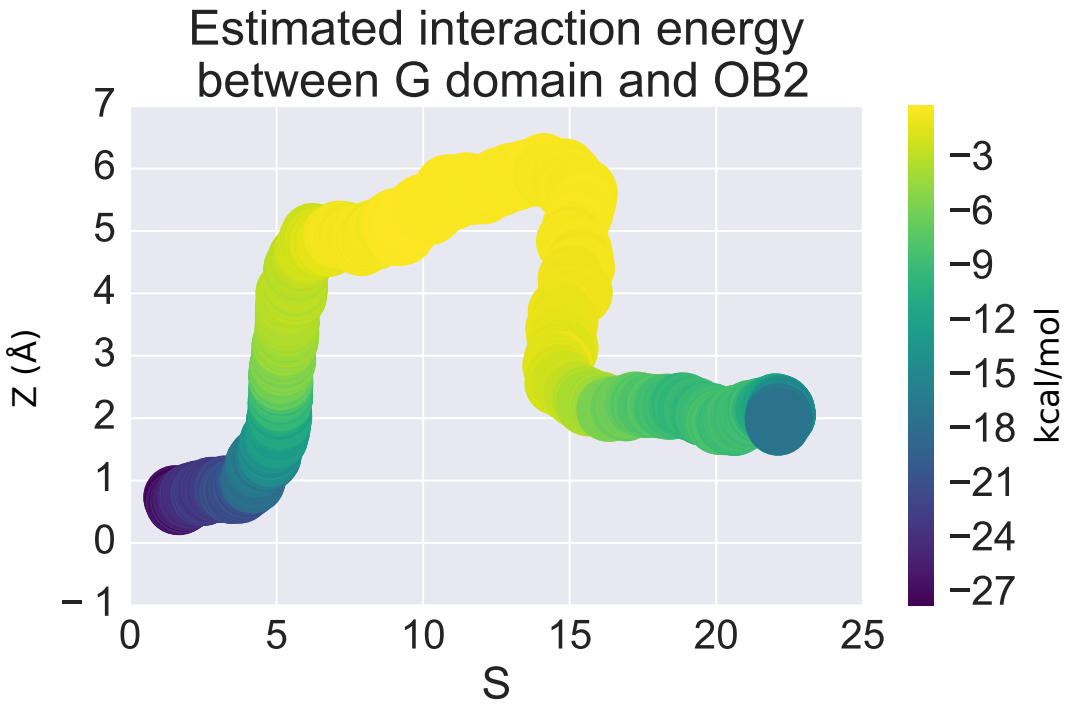

Figure 3.13: Interaction energies between the Domain 1 and Domain 3

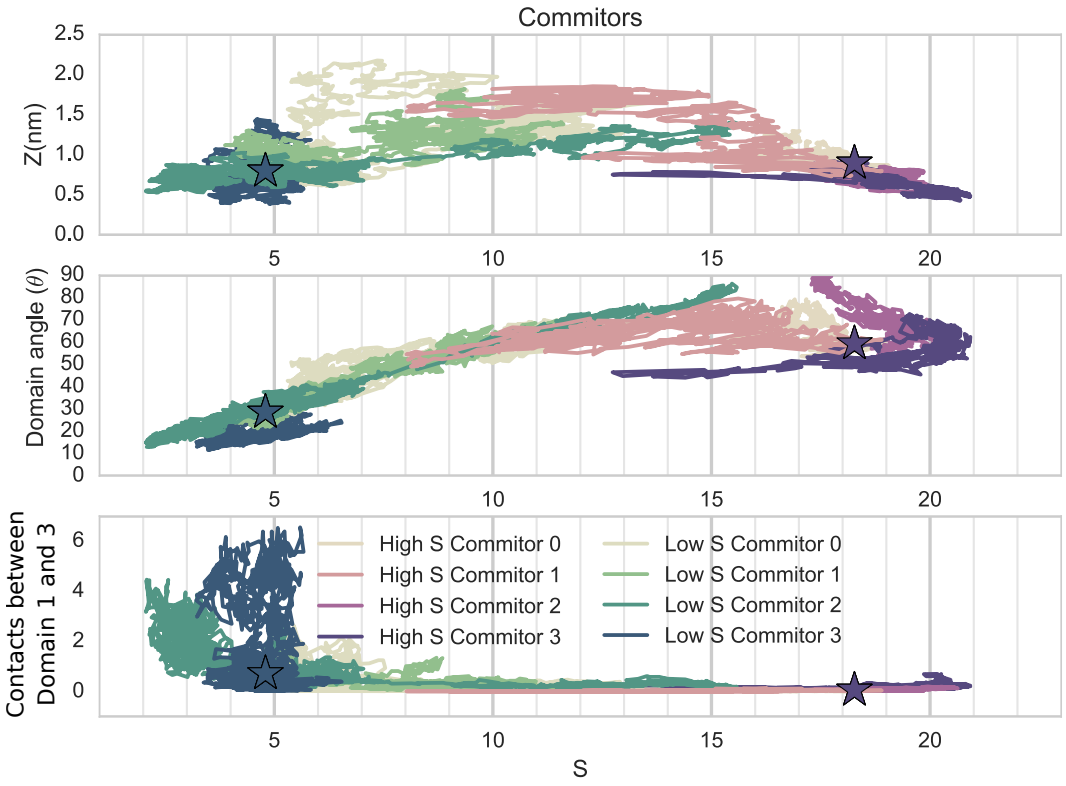

Figure 3.14: Unbiased MD simulations (commitors) from separation and rejoining states, indicated by stars. 50+ ns MD simulations.

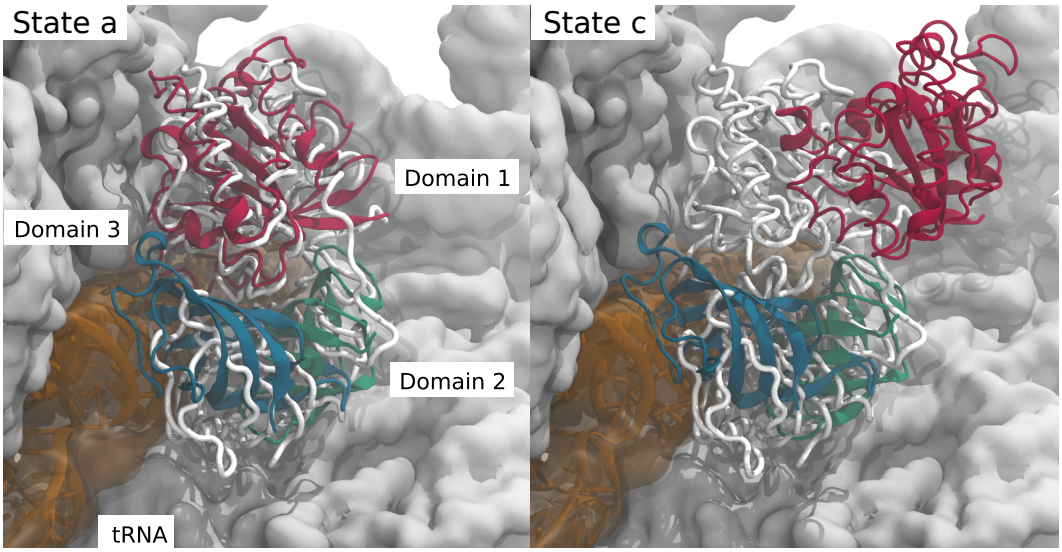

Figure 3.15: SMD trajectory of EF-Tu (red, green, and blue domains) with aa-tRNA (orange) aligned to ribosome PDB: 4V5G [78] (gray surface). Structures in state **a** easily fit within the ribosome without any clashes. As EF-Tu changes its conformation towards state **c**, Domain 1 (red) moves into the solution and away from the ribosome where it does not make any contacts. The pre-hydrolysis conformation of EF-Tu (white) and tRNA (orange, surface) in ribosome PDB: 4V5G was used for alignment. Of the residues interacting with the aa-tRNA identified in Eargle, et al [31], R300, R330, R339, and K376 maintain their interactions with aa-tRNA throughout the MFP. Alignment of trajectory and rendering performed using VMD [36].

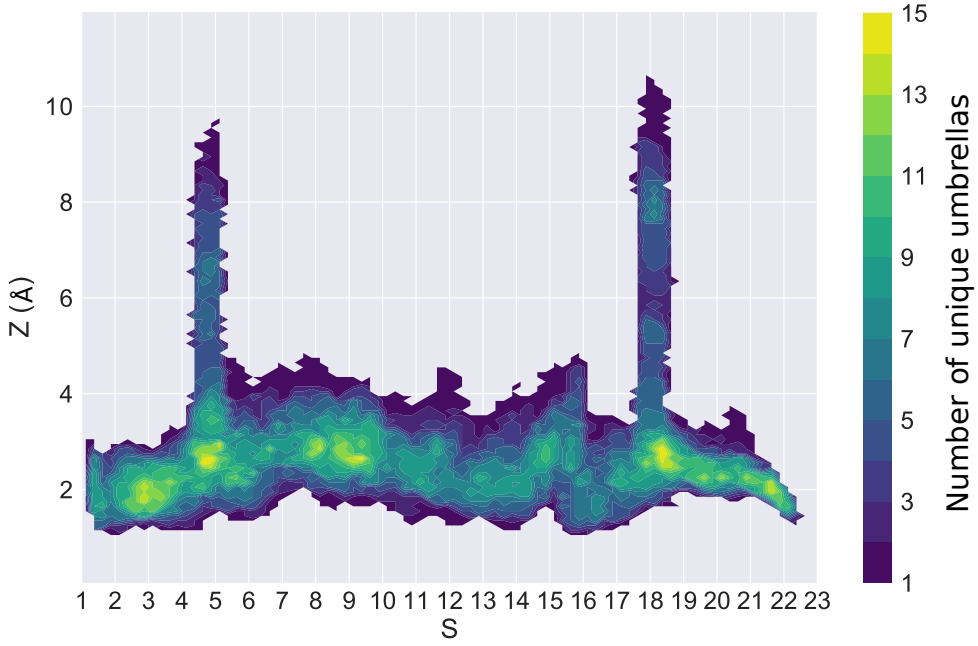

Figure 3.16: Overlap of 2-D umbrellas in S and Z space.

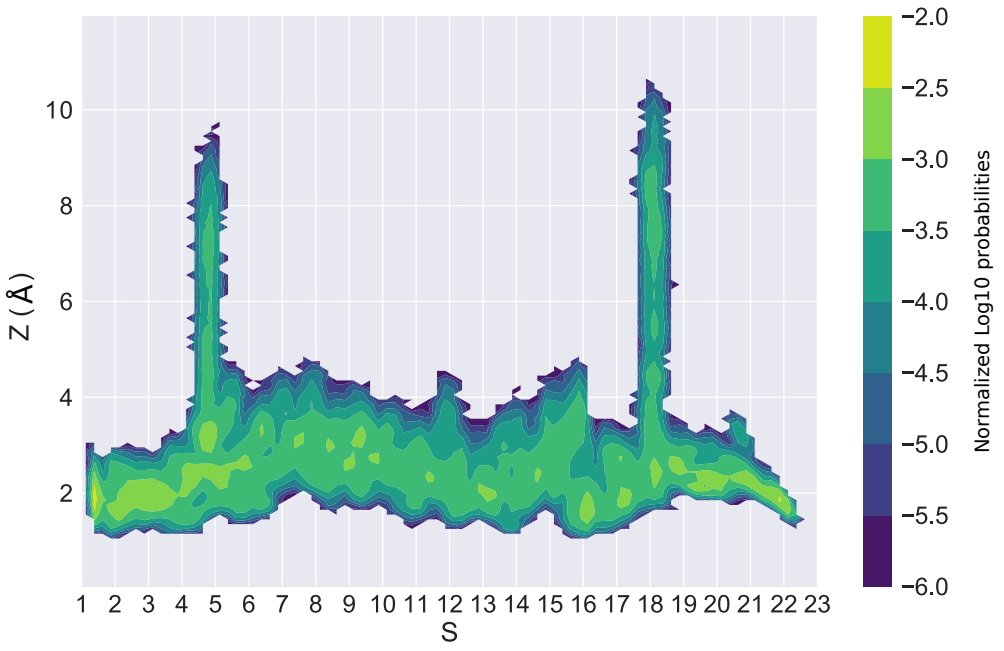

Figure 3.17: Probability density map of states visited by 2-D walkers. Plotted on Log10Norm colorscale.

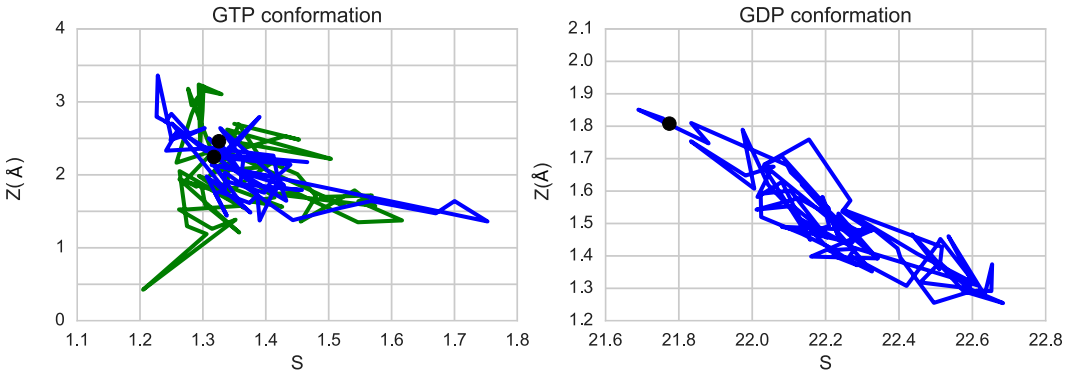

Figure 3.18: 100+ ns simulations of the GTP and GDP conformation (Commitors for endpoints).

Chapter 4

Pantoea metabolism[†]

4.1 Abstract

Pantoea sp. YR343, a rhizospheric bacteria, interacts with plants through the conversion of tryptophan (Trp) secreted from roots into a diverse spectrum of indole derivatives, such as indole-3-acetic acid (IAA), which promote plant growth. As a first step to a quantitative understanding of this symbiotic relationship, we have reconstructed the metabolic model of *Pantoea sp. YR343* starting from a well characterized model from a related species, *E. coli*. Using a combination of bioinformatics and proteomics data, the *E. coli* model was extended using gapfilling to include the reactions involved in Trp metabolism. Proteomics data and flux balance analysis (FBA) was used to predict the relative rates of secondary metabolite production, which was validated using mass spectrometry (MS). The majority of metabolites in the Trp catabolic pathway and their deuterated forms were identified based upon mass match and characteristic fragmentation patterns. A kinetic model of Trp metabolism was developed, allowing for the incorporation of

[†]Work includes contributions from Tyler Earnest, Piyush Labhsetwar, Bin Li, Jonathan Sweedler, Robert Standaert, Gregory Hurst, Jennifer Morrel-Falvey, Mitchel Doktyczm, and Zan Luthey-Schulten. Material from the Supporting Information can be found here: https://uofi.box.com/v/JonathanLaiThesis2017

the effect of competitive inhibition into the FBA model through corrections in the flux constraints. The improved model was then used to measure the sensitivity of IAA production on external Trp concentration at different glucose and oxygen levels, a relationship critical for quantitative modeling of the root/microbe interactions.

4.2 Introduction

The soil surrounding plant roots (rhizosphere) is a nutrient-rich environment that supports a diverse range of bacterial species. Many of these bacteria share a commensal or symbiotic relationship with the roots. The bacteria contribute to the relationship by converting local nutrients into a diverse range of metabolites that can both promote plant growth and regulate microbial populations [117–120] (Fig: 4.1). One such bacterial species, *Pantoea sp. YR343*, is well documented to import and catabolize exogenous Trp to produce plant auxins (e.g. IAA and tryptophol (TOL)). The catabolism of Trp can take place through multiple pathways [121]; however, the pathway which is used in *Pantoea sp. YR343* is not yet known.

One method to predict chemical secretion is through genome scale metabolic reconstruction [122]. Genome-scale metabolic reconstructions have been used to calculate growth rate and reaction utilization for organisms from all domains of life [123–126]. Briefly, the genome scale reconstructions represent the metabolic network of an organism as a interconnected set of chemical reactions catalyzed by enzymes encoded by the genome of the organism.

Flux through these chemical reactions are usually constrained by experimental data, e.g. proteomics, biochemical observations, and other assays, to restrict reactions to physiological minimal (V_{min}) or maximum (V_{max}) reaction fluxes [127] A biomass reaction is added to these networks to represent generation of metabolic components and energy for biomass production and enables prediction of growth rate in a given culturing medium. Using these metabolic networks to model steady state metabolism is premise of flux balance analysis (FBA) [**?**] which gives fluxes through all metabolic reactions to maximize an objective function usually the biomass reaction. Methods have been developed to automate the generation of metabolic reconstructions from annotated genomes—including *Pantoea sp. YR343* [128,129]. These tools use gene annotation to borrow metabolic reactions catalyzed by homologous proteins in related organisms to build draft reconstructions which needs gap filling and subsequent manual curation to be able to simulate growth. To quantitatively predict a bacterial secretome, however, involves imposing flux constraints on both the uptake/exchange reactions and internal reactions in the FBA model through the integration of –omics data [**?**].

In this paper, we constructed a genome scale FBA model of *Pantoea sp. YR343* and combined the results with a system of ordinary differential equations (ODEs) to predict the secondary metabolite secretion using proteomics data for *Pantoea sp. YR343* growing in M9 minimal media supplemented with 4% glucose and 1% Trp. We compared the diversity and concentration of metabolites secreted against intermediates detected by mass spectrometry (MS). The *Pantoea sp. YR343* FBA model is available in SBML format and all

of the analyses are available in a Jupyter notebook or SBML format.

4.3 Materials and Methods

4.3.1 Metabolic reconstruction

Gapfilling

We used the *E. coli iJO1366 model* [130] model as the basis of our metabolic reconstruction since, *Pantoea sp. YR343* is closely related to *E. coli* being in the same family of *Enterobacteriaceae*, sharing the same core metabolic network and primary metabolites. In total, more than half of the proteins are shared between *E. coli* and *Pantoea sp. YR343*. Additional reactions were added to the *Pantoea sp. YR343* model to account for secondary metabolites not found in *E. coli*. In order to add these missing enzymes and secondary metabolites, all possible degradation pathways for Trp were extracted from the KEGG database [131] as well as available literature. A schematic of all possible enzymes and metabolite intermediates identified is shown in (Fig: 4.2a). PSI-BLAST [132] profiles were constructed for each of the possible enzymes using proteins preferentially taken first from other *Pantoea* species, then *E. coli* and other Gamma-proteobacteria, and, finally, from any genome in the UniProt database. Best matches from the PSI-BLAST search to *Pantoea sp. YR343* and their e-value scores are given in 4.1. Because all of the PSI-BLAST queries are done with respect to the *Pantoea sp. YR343* genome (4696 proteins) and we are looking for proteins that carry out a specific enzymatic function, we

chose to adopt a conservative approach in determining an e-value threshold. Thus, we only considered a BLAST result a positive match if its e-value was less than 10^{-60}.

Implicit to this approach is the assumption that all reactions require enzymes to function. We acknowledge that several of the metabolic intermediates, especially the indole derivatives, might have spontaneous breakdown pathways under alkaline conditions [133, 134]. Under physiological conditions, however, Magnus and coworkers [135] argued against the spontaneous degradation of indole-3-pyruvate (IPA) and other indole derivatives. Thus, we adopted a conservative approach and assumed that all reactions capable of carrying flux require enzymes.

Determining flux constraints and simulating growth

Using the BLAST protocol above, we added five catabolism reactions, associated genes, enzymes, and metabolites (IPA, indole-3-lactate (ILA), indole-3-acetylaldehyde (IAAld), IAA, TOL), and transport reactions to the *Pantoea sp. YR343* model (4.1). Proteomics constraints were applied to these reactions in accordance with the protocol described in [136]. Briefly, the fluxes are constrained based on the enzyme turnover numbers taken from the BRENDA database [†] [137] and the number of proteins in *Pantoea sp. YR343* (4.5.3) and *E. coli* [138]. Glucose and Trp uptake rates (4.6 mmol $\cdot$ gdwt $^{-1}$ $\cdot$ hr $^{-1}$and 0.3 mmol $\cdot$ gdwt $^{-1}$ $\cdot$ hr $^{-1}$respectively) were determined assuming aerobic growth in M9 minimal medium supplemented with glucose (4mM, 4.5.2) and Trp

[†]Accessed on Aug. 31st, 2016

(1mM, 4.5.2). Components of the growth medium can be found in table (4.2). All other export reactions are unconstrained. A table of constants is given in the Supporting Information (4.3).

4.3.2 kcal·mol^{-1}

To study Trp catabolism as a function of time, we assumed that all of the enzymes in the catabolism pathway obeyed Michaelis-Menten kinetics and can be described through a system of ODEs.

Briefly, the ODEs are:

Elementary transport reactions

$$V^{max}_{Trp\ export} = V^{max}_{Trp\ uptake}$$

$$V^{transport}_{Trp} = \frac{k^{max}_{Trp\ uptake} * [Trp^{external}(aq)]}{k_m^{Trp\ uptake} + [Trp^{external}(aq)]}$$

$$V^{transport}_{\chi=IPA,IAAld,IAA,TOL} = \frac{\frac{1}{2}k^{max}_{Trp\ export} * [\chi(aq)]}{k_m^{\chi\ uptake} + [\chi(aq)]}$$

Elementary enzymatic reactions

$$k^{''enzyme"}_{max} = k^{''enzyme"}_{cat} * Number\_of\_enzyme$$

$$V_{Trp\rightarrow IPA} = \frac{k^{Tryptophan\ aminotransferase}_{max} * [Trp(aq)]}{k_m^{Tryptophan\ aminotransferase} + [Trp(aq)]}$$

$$V_{IPA\rightarrow IAAld} = \frac{k^{Indolepyruvate\ decarboxylase}_{max} * [IPA(aq)]}{k_m^{Indolepyruvate\ decarboxylase} + [IPA(aq)]}$$

$$V_{IAAld\rightarrow IAA} = \frac{k^{Aldehyde\ dehydrogenase}_{max} * [IAAld(aq)]}{k_m^{Aldehyde\ dehydrogenase} + [IAAld(aq)]}$$

$$V_{IAAld\rightarrow TOL} = \frac{k^{Alcohol\ dehydrogenase}_{max} * [IAAld](aq)}{k_m^{Alcohol\ dehydrogenase} + [IAAld(aq)]}$$

Time evolution equations for metabolites

$$\frac{d[Trp(aq)]}{dt} = V_{Trp}^{transport} - V_{Trp\rightarrow IPA}^{enzymatic}$$

$$\frac{d[IPA(aq)]}{dt} = -V_{IPA}^{transport} + V_{Trp\rightarrow IPA}^{enzymatic} - V_{IPA\rightarrow IAAld}^{enzymatic}$$

$$\frac{d[IAAld(aq)]}{dt} = -V_{IAALD}^{transport} + V_{IPA\rightarrow IAAld}^{enzymatic} - V_{IAald\rightarrow IAA}^{enzymatic} - V_{IAald\rightarrow TOL}^{enzymatic}$$

$$\frac{d[IAA(aq)]}{dt} = -V_{IAA}^{transport} + V_{IAald\rightarrow IAA}^{enzymatic}$$

$$\frac{d[TOL(aq)]}{dt} = -V_{TOL}^{transport} + V_{IAald\rightarrow TOL}^{enzymatic}$$

Trp and the other indole derivatives have a limited solublity in water; thus, if the concentration of these species exceeded the solubility threshold, $k_s^{species}$, the excess concentration would move into the "hydrophobic" phase of the cells. For example, if the concentration of [Trp (aq)] exceeds 65 mM, the difference would end up in the [Trp (s)] fraction until the [Trp (aq)] dropped below the solubility limit. In our implementation of the ODEs (available in the SI), we kept track of the soluble and insoluble fractions of indole metabolites. For clarity sake, however, we will only report on the total indole derivative concentrations.

4.3.3 Implementation

The km is integrated numerically using the LSODA algorithm from Scipy [139, 140] and the total error is ensured to be less than 10^{-3} mM. Parsimonious

FBA (pFBA) solutions are solved using the freely available COBRApy [141] libraries. An Jupyter [142] notebook (kinetic and FBA) and SBML (FBA only) file for running the *Pantoea sp. YR343* simulations with different media, as well as all of the required data, is freely available at www.scs.illinois.edu/schulten/software/index.html and can be found in the SI. Plots are generated using a Matplotlib [115] and Escher [143].

4.3.4 Metabolite Extraction from Pantoea YR343 Culture

Culture supernatant was acidified to pH 2.5 with 5N HCl and the metabolites were extracted twice with equal volumes of ethyl acetate. The ethyl acetate layers were pooled and evaporated to dryness under vacuum in rotary flash evaporator. After complete dryness, the residue was dissolved in 0.1 ml methanol:H2O (1:9, V/V), and used for direct infusion MS analysis.

4.3.5 Identification of IAA and its metabolites compounds

Direct infusion MS was performed using a maXis 4G Qq-ToF mass spectrometer (Bruker Daltonics) operated in positive ion mode. Molecular features were assigned with mass match and characteristic fragmentation patterns according to the literature [144, 145]

4.4 Results and Discussion

4.4.1 Metabolic model predicts experimentally measured growth rate

Using the bioinformatics approach described in the Methods section, we constructed a network of all possible reactions connecting Trp to all indole derivatives assumed to be secreted by *Pantoea sp. YR343* 4.1. If we assume a maximum glucose uptake rate of 4.6 mmol · gdwt $^{-1}$ · hr $^{-1}$, the model predicts that the cells will import glucose at a rate of 4.6 mmol · gdwt $^{-1}$ · hr $^{-1}$and that the cells will grow at a rate of 0.25^{-1} (2 hours 45 minutes doubling time). The computational growth rate of 0.25 compares favorably with the experimentally measured aerobic growth rate of 0.23 hr^{-1} (3 hour growth rates doubling time) under identical conditions reported elsewhere [**?**].

For the pFBA model, we set a maximum Trp uptake rate of 0.3 mmol · gdwt $^{-1}$ · hr $^{-1}$. The steady-state flux through the Trp transport reactions, however, never exceeded 0.09 mmol · gdwt $^{-1}$ · hr $^{-1}$, suggesting that the cell has spare Trp transport capacity. If we make the assumption that the indole-derivatives can be transported out of the cell through the same system that import Trp, we can safely say that the import and export reactions are probably not rate limiting during the catabolism process.

4.4.2 Kinetic Model of Tryptophan Catabolism

Using the constraints (Tab: 4.3) to describe the Trp catabolism, we modeled a 1 L well-stirred, liquid culture of *Pantoea sp. YR343* containing 5^{11} cells and 1 millimole of Trp. The solution to the kcal·mol^{-1} suggests that the cells produce various secondary metabolites after 96 hours, with IAA (0.08 mM), dominating the secretome (Fig: 4.3). The dominant species in the cell are IPA and IAA all other indole intermediates are in the micromolar concentrations. The concentration of IAA is directly proportional to the amount of Indolepyruvate decarboxylase (ipaD) in *Pantoea sp. YR343* (Fig: 4.4).

4.5 Conclusion

To our knowledge, this is the most comprehensive metabolic model of *Pantoea sp. YR343*, containing both proteomic data and possible candidate genes. The model connects reactions from the ipaD pathway to central metabolism in *E. coli*. Both the kinetic and pFBA model accurately reproduce the secondary metabolite production under aerobic conditions as determined by MS experiments.

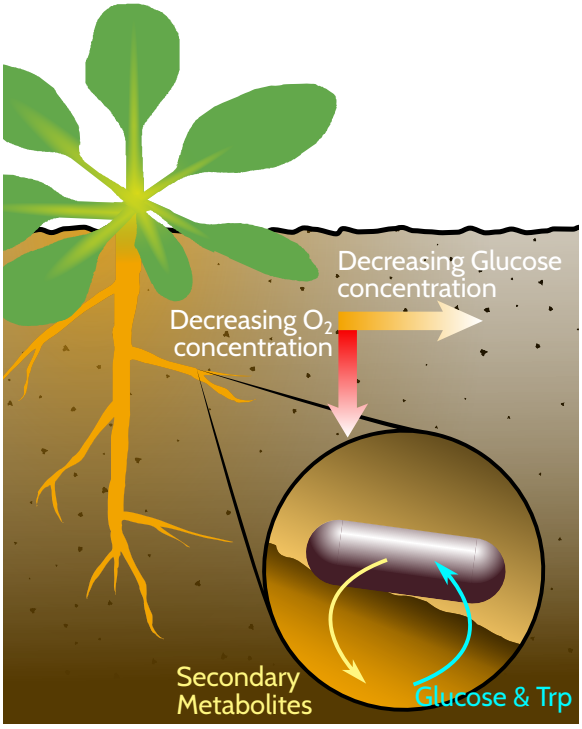

Figure 4.1: Schematic of a rhizosphere bacteria (e.g. *Pantoea sp. YR343*) in its native environment.

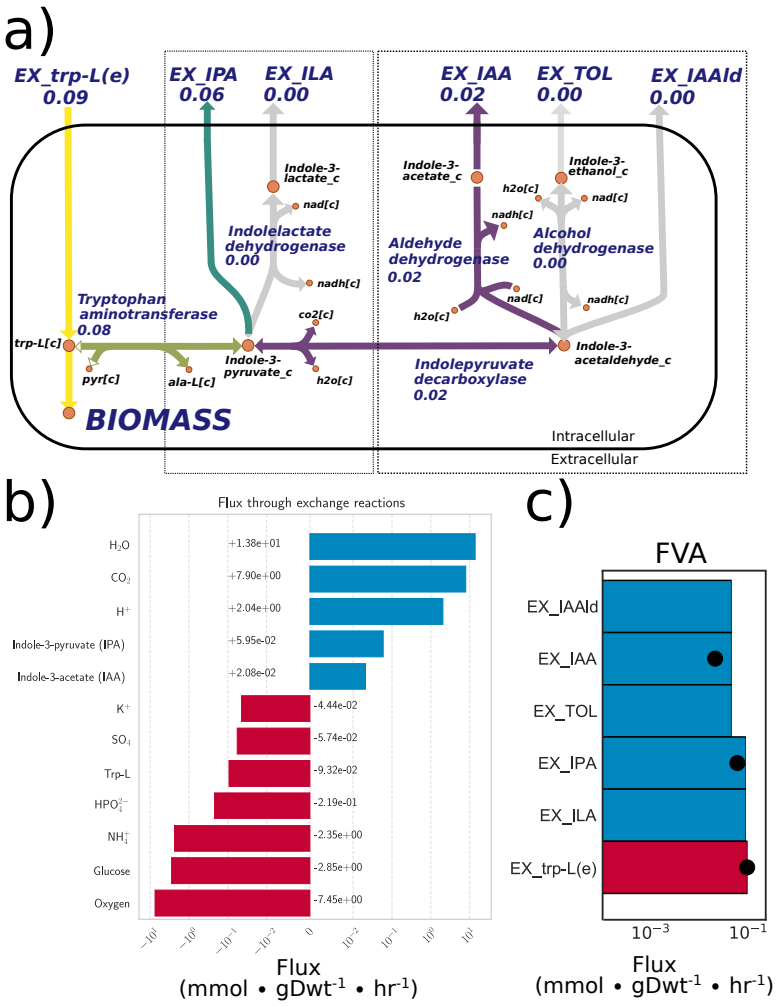

Figure 4.2: A metabolic network for metabolism in *Pantoea sp. YR343*. (A) Network of reactions for Trp catabolism in *Pantoea sp. YR343*. Metabolites are shown in circles while reactions are shown as edges. Edge color denotes the flux through each of the reaction edges and the color ranges from gray (0 mmol · gdwt $^{-1}$ · hr $^{-1}$) to purple to yellow (>0.09 mmol · gdwt $^{-1}$ · hr $^{-1}$), Flux through the network is calculated assuming a maximum glucose/Trp uptake rate of 4.57 and 0.3 mmol · gdwt $^{-1}$ · hr $^{-1}$ respectively. This yields a growth rate of 0.25 hr^{-1} and a doubling time of 2.8 hours. (B) Import (red) and export (blue) of metabolites from pFBA solution. (C) Variation in metabolite export fluxes. The variation in the export of indole derivatives is growth rate independent.

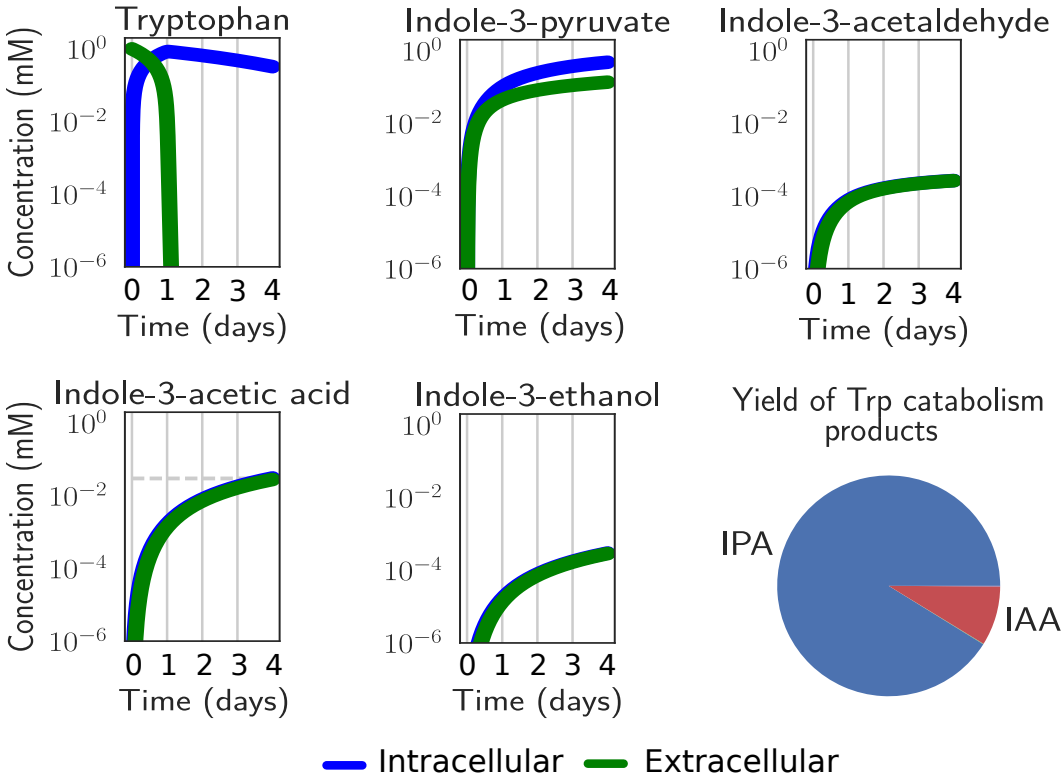

Figure 4.3: System of ODEs describing Trp catabolism. ODEs assumes a colony of *Pantoea sp. YR343* with $5 * 10^{11}$ cells per liter growing in the *in silico* M9 media 4.2. Dashed line in the IAA panels indicate the experimentally measured IAA concentration (0.03 mM) from the mass spec and proteomics experiments and it favorably compares to the predicted ODE value after 4 days (0.08 mM).

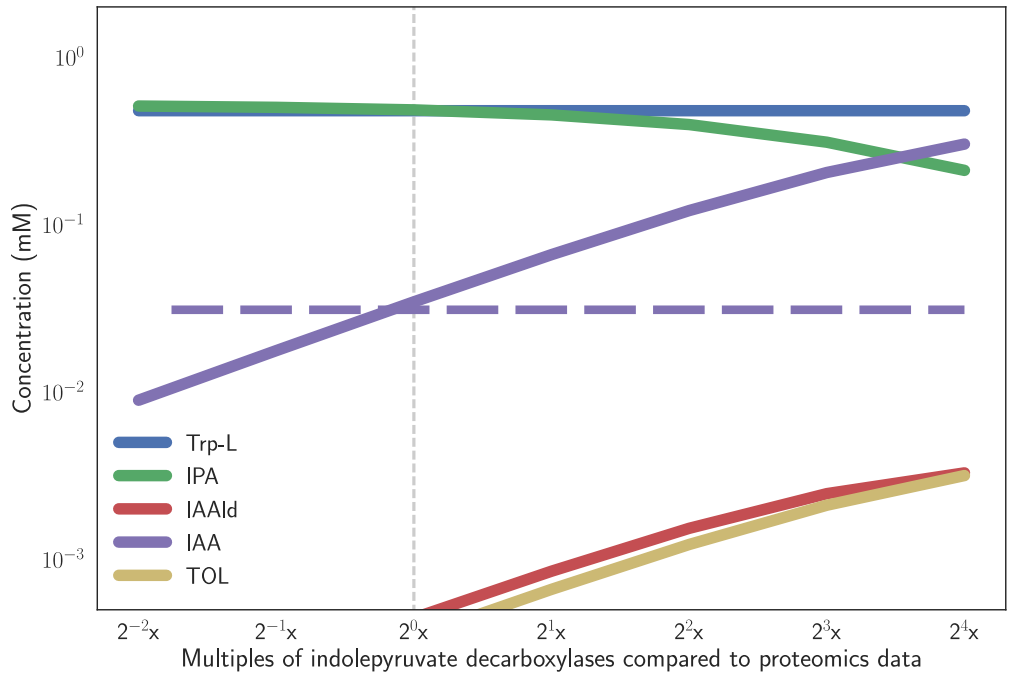

Figure 4.4: Change in IAA and derivative production as a function of the number of ipaD enzymes in *Pantoea sp. YR343*. Dashed line shows the experimentally measured IAA concentration (0.03 mM) from the mass spec and proteomics experiments.

4.5.1 Scripts and FBA model

All of the Python scripts and data used to model metabolism are freely available here: www.scs.illinois.edu/schulten/software/index.html. Initial model for *E. coli* used to build *Pantoea sp. YR343* given in the MATLAB file: *populationFBA_ZLS_Data.mat*. The example Python notebook scans various glucose, tryptophan, and oxygen concentrations. The model is also available in SBML format. You can directly download the Python notebook here:

https://uofi.box.com/s/6zjbdj8ssqe6ugbh71hvd9l875oqnsw4

or through the main URL: **https://uofi.box.com/v/JonathanLaiThesis2017**.

4.5.2 Determining uptake rates for glucose and Trp

Glucose uptake

Based on the proteomics data, there are approximate 2.7 $\times$ 10^6 proteins in the cell (Tab: 4.5.3)—of which, approximate 1020 proteins are identified as PtsG complex IIBC transporters. Assuming an average turnover rate of 210s^{-1} **BIONUMBER: 103693** and a cell dry weight of 280fg [146], the total uptake rate of glucose is:

$$\begin{aligned} glucose\ uptake =& \frac{[Turnover] \times [Protein\ count] \times [3600s \cdot hr^{-1}]}{Avogadro's\ number \times dry\ weight\ cell} \\ =& \frac{210s^{-1} \times 1020 \times 3600s \cdot hr^{-1}}{6.022 \times 10^{20} mmoles^{-1} \times 280 \times 10^{-15} g\_dry\_weight} \\ =& \mathbf{4.57 mmoles \cdot hr^{-1} \cdot g\_dry\_weight^{-1}} \end{aligned}$$

Trp uptake

According to [PMID: 4880290], the maximum flux for Trp uptake is:

$$Vmax = 0.59 mmoles \cdot 30 - sec^{-1} \cdot kg\_wet\,weight^{-1} \qquad [PMID: 4880290]$$

Convert velocity to per hour

$$Vmax = 0.0196 mmoles \cdot s^{-1} \cdot kg\_wet\_weight^{-1} \times \frac{1\,30 - sec}{30s} \times \frac{3600s}{1hr} = 70.8$$

Convert velocity to per g_{wet_weight}

$$Vmax = 70.8 mmoles \cdot hr^{-1} \cdot kg\_wet\_weight^{-1} \times \frac{1kg\_wet\_weight}{1000g\_wet\_weight} = 0.0708$$

Convert velocity from wet weight to dry weight From Bionumbers 109836, the ratio of wet to dry weight is:

$$\frac{Wet}{Dry} = \frac{1.7g \cdot L^{-1}}{0.39g \cdot L^{-1}} = 4.36 g\_wet\_weight \cdot g\_dry\_weight^{-1}$$

The new Vmax is:

$$Vmax = 0.0708 mmoles \times \frac{Wet}{Dry}$$

$$Vmax = 0.0708 mmoles \cdot hr^{-1} \times 4.36 \cdot g\_dry\_weight^{-1} = 0.309$$

Vmax for Trp transport is:

$$Vmax = 0.309 mmoles \cdot hr^{-1} \cdot g\_dry\_weight^{-1}$$

Assuming a buffer with 200 mg of Trp per liter and a Trp molecular weight: 204.225 g/mol = 204.225 mg/mmol

$$[S] = \frac{200mg \cdot L^{-1}}{204.225mg \cdot mmol^{-1}} [S] = 0.979 mmol^{1} \cdot L^{-1} = 0.979 mM$$

Given that $Km \equiv 0.9uM = 0.0009mM[PMID: 4880290]$ and $[S] >>> k_m$, then:

$$\textbf{Vmax = 0.309 mmol} \cdot \textbf{gdwt}^{-1} \cdot \textbf{hr}^{-1}$$

4.5.3 Protein counts

Proteomics experiments of *Pantoea sp. YR343* were performed in triplicate in M9 minimial media supplemented with 0.4% glucose and 0.1% tryptophan. Values from the proteomics study were reported as NSAF which is directly proportional to the to the absolute number of proteins in the cell. The absolute protein counts were calculated by multiplying the median NSAF value by the total number of proteins in the cell (assumed to be approximately 4×10^6 proteins given a 1 fL cell [147]). An Excel spreadsheet with all of the data used to convert the proteomic data to absolute protein counts can be found here: **https://uofi.box.com/s/l7bre05mpnmtzi01724is7te7rc66m56** or through the main URL: **https://uofi.box.com/v/JonathanLaiThesis2017**.

Table 4.1: Tryptophan catabolism reactions

Protein name	Reactants and Products in Reaction	EC Num. [131]	Uniprot ID	Blast eValue score
Tryptophan aminotransferase	Trp + Pyruvate = Alanine + IPA	(2.6.1.27)	J3CG65	0
Indolepyruvate decarboxylase	IPA = CO_2 + IAAld	(4.1.1.74)	J2VY20	$2 * 10^{-152}$
Aldehyde dehydrogenase	IAAld + NAD^+ + H_2O = NADH + IAA	(1.2.1.3)	J2VCM3	0
Alcohol dehydrogenase	IAAld + NADH = NAD^+ + H_2O + TOL	(1.1.1.1)	J2VNB9	$9 * 10^{-42}$

Table 4.2: *In silico* M9 media

Components of the *in silico* M9 media supplemented with glucose and Trp

Component	Flux upper bound (mmol · gdwt $^{-1}$ · hr $^{-1}$)
M9 salts	
Cl^-	1000
K^+	1000
Na^+	1000
NH_4^+	1000
PO_4^{3-}	1000
Trace metals	
Ca^{2+}	1000
Co	1000
Cu^+	1000
Fe^{2+}	1000
Mn^{2+}	1000
MoO_4^{2-}	1000
Mg^{2+}	1000
Ni^{2+}	1000
Se	1000
SO_4^{2-}	1000
W	1000
Zn^{2+}	1000
Supplemental components	
Glucose	4.3
Oxygen	20.0
Tryptophan	0.3
Water	1000

Table 4.3: Reaction and kinetic parameters used in indole-catabolism model

Protein name	k_{cat} (s^{-1})	k_M (mM)	Species	Protein counts
Tryptophan aminotransferase	19.40† [148]	0.29 [148]	*Arabidopsis thaliana*	198
Indolepyruvate decarboxylase	2 [664109]	0.79 [651291]	*Sulfolobus sp.*	488
Aldehyde dehydrogenase	26 [690591]	0.02 [348694]	*E. coli*	1067
Alcohol dehydrogenase	1.83 [285569]	0.2 [285569]	*H. sapiens*	13684

Table 4.4: Intermediates detected by mass spec

Identification	Theoretical mass	Measured mass [M+H]+	Mass accuracy (ppm)	Fragment ions (m/z)
Indole-3-acetic acid (IAA)	176.0706	176.0717	6.2	130.1
Indole-3-pyruvic acid (IPA) and decomposition products				
IPA directly detected	204.0655	204.0665	4.9	144
Indole	118.0651	118.0659	6.8	91.1, 101.0
Indole-3-carboxaldehyde (ICA)	146.06	146.0609	5	91.1, 118.1

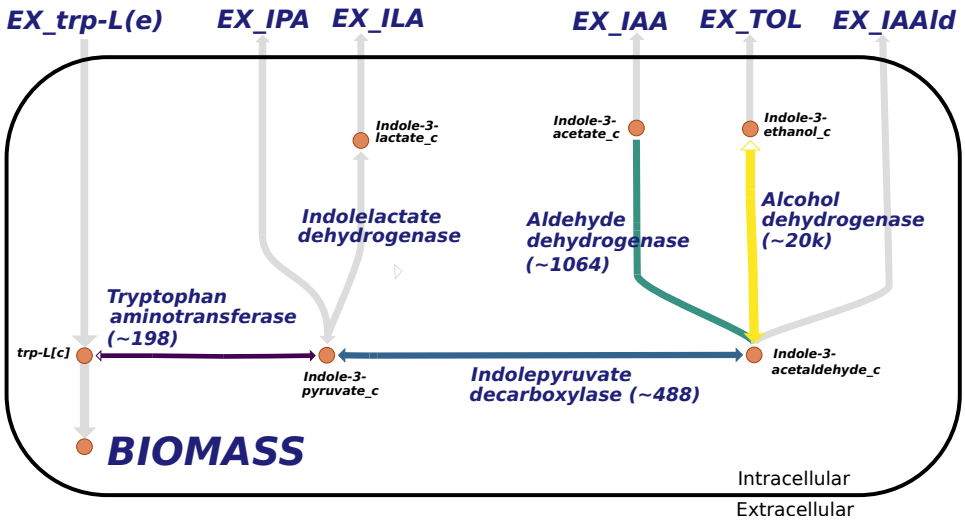

Figure 4.5: Trp catabolism reactions added to the *Pantoea sp. YR343* network and the estimated protein count.

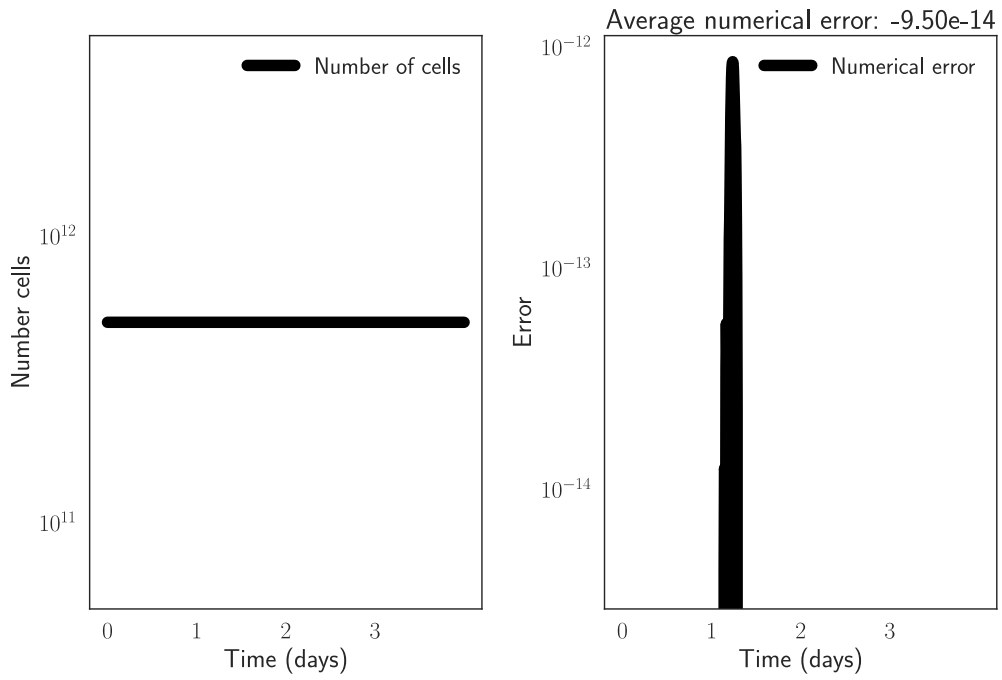

Figure 4.6: Numerical stability of the ODEs. Left panel shows the number of cells in the system while right panel shows the change in the total number of metabolites in the ODEs.

References

[1] Jonathan Lai, Ke Chen, and Zaida Luthey-Schulten. Structural Intermediates and Folding Events in the Early Assembly of the Ribosomal Small Subunit. 42:13335–13345, 2013.

[2] Ke Chen, John Eargle, Jonathan Lai, Hajin Kim, Sanjaya Abeysirigunawardena, Megan Mayerle, Sarah Woodson, Taekjip Ha, and Zaida Luthey-Schulten. Assembly of the Five-way Junction in the Ribosomal Small Subunit using Hybrid MD-Gō Simulations. 116(23):6819–31, June 2012.

[3] W.A. Held, B. Ballou, S. Mizushima, and M. Nomura. Assembly Mapping of 30S Ribosomal Proteins from *E. coli*: Further studies. 249(10):3103–3111, 1974.

[4] T. Adilakshmi, P. Ramaswamy, and S.A. Woodson. Protein-Independent Folding Pathway of the 16S rRNA 5′ domain. 351(3):508–519, August 2005.

[5] Tadepalli Adilakshmi, Deepti L Bellur, and Sarah a Woodson. Concurrent Nucleation of 16S Folding and Induced Fit in 30S Ribosome Assembly. 455(7217):1268–72, October 2008.

[6] M.W.T. Talkington, G. Siuzdak, and J.R. Williamson. An Assembly Landscape for the 30S Ribosomal Subunit. 438(7068):628–632, 2005.

[7] M.T. Sykes and J.R. Williamson. A Complex Assembly Landscape for the 30S Ribosomal Subunit. 38:197–215, 2009.

[8] A.M. Mulder, C. Yoshioka, A.H. Beck, A.E. Bunner, R.A. Milligan, C.S. Potter, B. Carragher, and J.R. Williamson. Visualizing Ribosome Biogenesis: Parallel Assembly Pathways for the 30S Subunit. 330(6004):673–677, October 2010.

[9] Keith Connolly and Gloria Culver. Deconstructing Ribosome Construction. 34(5):256–263, 2009.

[10] CJ Weitzmann, PR Cunningham, K Nurse, and J. Ofengand. Chemical evidence for domain assembly of the *E. coli* 30S ribosome. 7(1):177–180, 1993.

[11] Priya Ramaswamy and Sarah A Woodson. Global Stabilization of rRNA Structure by Ribosomal Proteins S4, S17, and S20. 392(3):666–677, September 2009.

[12] Thomas Becker, Shashi Bhushan, Alexander Jarasch, Jean-Paul Armache, Soledad Funes, Fabrice Jossinet, James Gumbart, Thorsten Mielke, Otto Berninghausen, Klaus Schulten, Eric Westhof, Reid Gilmore, Elisabet C Mandon, and Roland Beckmann. Structure of Monomeric Yeast and Mammalian Sec61 Complexes Interacting with the Translating Ribosome. 326(5958):1369–1373, December 2009.

[13] Birgit Seidelt, C Axel Innis, Daniel N Wilson, Marco Gartmann, Jean-Paul Armache, Elizabeth Villa, Leonardo G Trabuco, Thomas Becker, Thorsten Mielke, Klaus Schulten, Thomas a Steitz, and Roland Beckmann. Structural Insight into Nascent Polypeptide Chain-Mediated Translational Stalling. 326(5958):1412–1415, December 2009.

[14] Leonardo G Trabuco, Eduard Schreiner, John Eargle, Peter Cornish, Taekjip Ha, Zaida Luthey-Schulten, and Klaus Schulten. The Role of L1 Stalk-tRNA Interaction in the Ribosome Elongation Cycle. 402(4):741–760, October 2010.

[15] Jean-Paul Armache, Alexander Jarasch, Andreas M Anger, Elizabeth Villa, Thomas Becker, Shashi Bhushan, Fabrice Jossinet, Michael Habeck, Gülcin Dindar, Sibylle Franckenberg, Viter Marquez, Thorsten Mielke, Michael Thomm, Otto Berninghausen, Birgitta Beatrix, Johannes Söding, Eric Westhof, Daniel N Wilson, and Roland Beckmann. Cryo-EM Structure and rRNA Model of a Translating Eukaryotic 80S Ribosome at 5.5-Å Resolution. 107(46):19748–19753, November 2010.

[16] Paul C Whitford, Peter Geggier, Roger B Altman, Scott C Blanchard, José N Onuchic, and Karissa Y Sanbonmatsu. Accommodation of Aminoacyl-tRNA into the Ribosome Involves Reversible Excursions Along Multiple Pathways. 16(6):1196–1204, June 2010.

[17] S.M. Stagg, A. Mears, and Harvey S.C. A Structural Model for the Assembly of the 30S Subunit of the Ribosome. 328:49–61, 2003.

[18] J. Trylska, J.A. McCammon, and C.L. Brooks III. Exploring Assembly Energetics of the 30S Ribosomal Subunit Using an Implicit Solvent Approach. 127(31):11125–11133, 2005.

[19] Qizhi Cui, Robert K. Z. Tan, Stephen C. Harvey, and David a. Case. Low-Resolution Molecular Dynamics Simulations of the 30S Ribosomal Subunit. 5(4):1248–1263, January 2006.

[20] K. Hamacher, J. Trylska, and J.A. McCammon. Dependency Map of Proteins in the Small Ribosomal Subunit. 2(2):e10, February 2006.

[21] Brittany Burton, Michael T. Zimmermann, Robert L. Jernigan, and Yongmei Wang. A computational investigation on the connection between dynamics properties of ribosomal proteins and ribosome assembly. 8(5):e1002530, 05 2012.

[22] D.L. Bellur and S.A. Woodson. A Minimized rRNA-Binding Site for Ribosomal Protein S4 and its Implications for 30S Assembly. 37(6):1886–1896, 2009.

[23] K. Chen, J. Eargle, K. Sarkar, M. Gruebele, and Z. Luthey-Schulten. Functional Role of Ribosomal Signatures. 99(12):3930 – 3940, 2010.

[24] B.A. Shoemaker, J.J. Portman, and P.G. Wolynes. Speeding Molecular Recognition by Using the Folding Funnel: the Fly-Casting Mechanism. 97(16):8868–8873, 2000.

[25] Y. Levy, J.N. Onuchic, and P.G. Wolynes. Fly-Casting in Protein-DNA Binding: Frustration Between Protein Folding and Electrostatics Facilitates Target Recognition. 129(4):738–739, 2007.

[26] Y. Huang and Z. Liu. Kinetic Advantage of Intrinsically Disordered Proteins in Coupled Folding-Binding Process: a Critical Assessment of the "Fly-Casting" Mechanism. 393(5):1143–1159, 2009.

[27] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. Theory of Protein Folding: the Energy Landscape Perspective. 48(1):545–600, 1997.

[28] Veysel Berk, Wen Zhang, Raj D. Pai, and Jamie H. D Cate. Structural Basis for mRNA and tRNA Positioning on the Ribosome. 103(43):15830–15834, 2006.

[29] Alexander D. MacKerell Jr., Michael Feig, and Charles L. Brooks III. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. pages 1400–1415, 2004.

[30] N. Foloppe and A. D. MacKerell Jr. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. 21:86–104, 2000.

[31] John Eargle, Alexis A. Black, Anurag Sethi, Leonardo G. Trabuco, and Zaida Luthey-Schulten. Dynamics of Recognition Between tRNA and Elongation Factor Tu. 377(5):1382 – 1405, 2008.

[32] John Eargle and Zaida Luthey-Schulten. Simulating dynamics in rna and protein complexes. In Neocles Leontis and Eric Westhof, editors, *RNA 3D Structure Analysis and Prediction*, volume 27 of *Nucleic Acids and Molecular Biology*, pages 213–238. Springer Berlin Heidelberg, 2012.

[33] Alexander Balaeff, John Eargle, and Elijah Roberts. *Ionize v 1.6*. University of Illinois – Urbana-Champaign, Urbana, Illinois, Website: `http://www.scs.illinois.edu/schulten/software/mdtools/ionize/`, 2005.

[34] JH Roh, RM Briber, A Damjanovic, D Thirumalai, SA Woodson, and AP Sokolov. Dynamics of tRNA at Different Levels of Hydration. 96(7):2755–2762, 2009.

[35] Helmut Grubmueller and Volker Groll. *Solvate v 1.0*. Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, Website: `http://www.mpibpc.mpg.de/home/grubmueller/downloads/solvate/index.html`, 1996.

[36] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD–Visual Molecular Dynamics. 14:33–38, 1996.

[37] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable Molecular Dynamics with NAMD. 26(16):1781–1802, Dec 2005.

[38] IN Serdyuk, ZV Gogia, S.Y. Venyaminov, NN Khechinashvili, VN Bushuev, and AS Spirin. Compact globular conformation of protein S4 from *Escherichia coli* ribosomes. *JMB*, 137(1):93–107, 1980.

[39] E.W. Sayers, R.B. Gerstner, D.E. Draper, and D.A. Torchia. Structural preordering in the N-terminal region of ribosomal protein S4 revealed by heteronuclear NMR spectroscopy. *BCH*, 39(44):13602–13613, 2000.

[40] Paul C. Whitford, Jeffrey K. Noel, Shachi Gosavi, Alexander Schug, Kevin Y. Sanbonmatsu, and José N. Onuchic. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. 75(2):430–441, 2009.

[41] Paul C. Whitford, Alexander Schug, John Saunders, Scott P. Hennelly, José N. Onuchic, and Kevin Y. Sanbonmatsu. Nonlocal helix formation is key to understanding s-adenosylmethionine-1 riboswitch function. 96(2):L7 – L9, 2009.

[42] Jeffrey K Noel, Paul C Whitford, Karissa Y Sanbonmatsu, and José N Onuchic. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. 38(Web Server issue):W657–61, July 2010.

[43] Sander Pronk, Szilàrd Pàll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. 2013.

[44] Jamie Cannone, Sankar Subramanian, Murray Schnare, James Collett, Lisa D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi Madabusi, Kirsten Muller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin Gutell. The Comparative RNA Web (CRW) Site: an Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs. 3(1):15, 2002.

[45] Anurag Sethi, John Eargle, Alexis A Black, and Zaida Luthey-Schulten. Dynamical Networks in tRNA: Protein Complexes. 106(16):6620–6625, 2009.

[46] Rebecca Alexander, John Eargle, and Zaida Luthey-Schulten. Experimental and computational determination of tRNA dynamics. 584(2):376–386, 2010.

[47] J. Eargle and Z. Luthey-Schulten. NetworkView: 3D Display and Analysis of Dynamic Structure Networks. 28:3000–3001, 2012.

[48] Samuel S. Cho, Yaakov Levy, and Peter G. Wolynes. P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. 103(3):586–591, 2006.

[49] Elijah Roberts, John Eargle, Dan Wright, and Zaida Luthey-Schulten. MultiSeq: unifying sequence and structure data for evolutionary analysis. 7:382, Aug 2006.

[50] Anne E Bunner, Stefan Nord, P Mikael Wikström, and James R Williamson. The Effect of Ribosome Assembly Cofactors on *in vitro* 30S Subunit Reconstitution. 398(1):1–7, May 2010.

[51] J. Eargle and Z. A. Luthey-Schulten. *RNA 3D Structure Analysis and Prediction*, chapter 12, pages 213–238. Springer, RNA 3D structure analysis and prediction edition, 2012.

[52] U Vogel and K F Jensen. The RNA chain elongation rate in Escherichia coli depends on the growth rate. 176(10):2807–2813, 1994.

[53] Ranjani Narayanan, Yogambigai Velmurugu, Serguei V. Kuznetsov, and Anjum Ansari. Fast folding of rna pseudoknots initiated by laser temperature-jump. 133(46):18767–18774, 2011.

[54] Konstantinos Paterakis, Jennifer Littlechild, and Paul Woolley. Structural and functional studies on protein s20 from the 30-s subunit of the escherichia coli ribosome. 129(3):543–548, 1983.

[55] Krishna Neupane, Dustin B. Ritchie, Hao Yu, Daniel A. N. Foster, Feng Wang, and Michael T. Woodside. Transition path times for nucleic acid folding determined from energy-landscape analysis of single-molecule trajectories. 109:068102, Aug 2012.

[56] R.R. Gutell. Comparative RNA Web Site and Project, 2002. http://www.rna.ccbb.utexas.edu/.

[57] E. Roberts, A. Sethi, J. Montoya, C.R. Woese, and Z. Luthey-Schulten. Molecular Signatures of Ribosomal Evolution. 105(37):13953–13958, 2008.

[58] Tyler M. Earnest, Jonathan Lai, Ke Chen, Michael J. Hallock, James R. Williamson, and Zaida Luthey-Schulten. Toward a whole-cell model of ribosome biogenesis: Kinetic modeling of SSU assembly. *Biophysical Journal*, 109(6):1117–1135, sep 2015.

[59] Sanjaya C. Abeysirigunawardena, Hajin Kim, Jonathan Lai, Kaushik Ragunathan, Zaida Luthey-Schulten, Taekjip Ha, and Sarah A. Woodson1. Assembly hierarchy of ribosomal complexes depends on protein-coupled rna dynamics. *Nature Communications*.

[60] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. *J. Chem. Theory Comput.*, 8(9):3257–3273, 2012.

[61] Elizabeth J. Denning, U. Deva Priyakumar, Lennart Nilsson, and Alexander D. Mackerell. Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *J. Chem. Theory Comput.*, 32(9):1929–1943, 2011.

[62] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. 79:926–935, 1983.

[63] P Auffinger and E Westhof. RNA hydration: three nanoseconds of multiple molecular dynamics simulations of the solvated tRNA(Asp) anticodon hairpin. 269(3):326–341, Jun 1997.

[64] G. Caliskan, C. Hyeon, U. Perez-Salas, R. M. Briber, S. A. Woodson, and D. Thirumalai. Persistence Length Changes Dramatically as RNA Folds. 95:268303, Dec 2005.

[65] Tyler M. Earnest, John A. Cole, Joseph R. Peterson, Michael J. Hallock, Thomas E. Kuhlman, and Zaida Luthey-Schulten. Ribosome biogenesis in replicating cells: Integration of experiment and theory. *Biopolymers*, 105(10):735–751, 2016.

[66] Hajin Kim, Sanjaya C Abeysirigunawarden, Ke Chen, Megan Mayerle, Kaushik Ragunathan, Zaida Luthey-Schulten, Taekjip Ha, and Sarah A Woodson. Protein-guided RNA dynamics during early ribosome assembly. *Nature*, 506(7488):334–338, 2014.

[67] R C Thompson, D B Dix, and A M Karim. The Reaction of Ribosomes with Elongation Factor Tu.GTP complexes. Aminoacyl-tRNA-independent Reactions in the Elongation cycle determine the Accuracy of Protein Synthesis. 261(11):4868–4874, 1986.

[68] Ute Kothe and Marina V. Rodnina. Delayed Release of Inorganic Phosphate from Elongation Factor Tu Following GTP Hydrolysis on the Ribosome. 45(42):12767–12774, Oct 2006.

[69] Lee E. Sanderson and Olke C. Uhlenbeck. Exploring the Specificity of Bacterial Elongation Factor Tu for Different tRNAs. 46(21):6194–6200, 2007. PMID: 17489561.

[70] A Pingoud, W Block, A Wittinghofer, H Wolf, and E Fischer. The elongation factor Tu binds aminoacyl-tRNA in the presence of GDP. 257(19):11261–7, 1982.

[71] Hans Bremer and Patrick P. Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1), sep 2008.

[72] Poul Nissen, Søren Thirup, Morten Kjeldgaard, and Jens Nyborg. The crystal structure of cys-tRNACys–EF-tu–GDPNP reveals general and specific features in the ternary complex and in tRNA. *Structure*, 7(2):143–156, feb 1999.

[73] Galina Polekhina, Søren Thirup, Morten Kjeldgaard, Poul Nissen, Corinna Lippmann, and Jens Nyborg. Helix unwinding in the effector region of elongation factor EF-Tu-GDP. *Structure*, 4(10):1141–1151, 1996.

[74] Tillmann Pape, Wolfgang Wintermeyer, and Marina V. Rodnina. Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E.coli ribosome. 17(24):7490–7497, 1998.

[75] Cristina Maracci and Marina V. Rodnina. Review: Translational GTPases. *Biopolymers*, 105(8):463–475, may 2016.

[76] MV Rodnina, R Fricke, L Kuhn, and W Wintermeyer. Codon-dependent conformational change of elongation factor Tu preceding GTP hydrolysis on the ribosome. 14(11):2613, 1995.

[77] Ka-Weng Ieong, Ülkü Uzun, Maria Selmer, and Måns Ehrenberg. Two proofreading steps amplify the accuracy of genetic code translation. 113(48):13744–13749, 2016.

[78] T. M. Schmeing, R. M. Voorhees, A. C. Kelley, Y.-G. Gao, F. V. Murphy, J. R. Weir, and V. Ramakrishnan. The crystal structure of the ribosome bound to EF-tu and aminoacyl-tRNA. *Science*, 326(5953):688–694, oct 2009.

[79] Xabier Agirrezabala, Eduard Schreiner, Leonardo G Trabuco, Jianlin Lei, Rodrigo F Ortiz-Meoz, Klaus Schulten, Rachel Green, and Joachim Frank. Structural insights into cognate versus near-cognate discrimination during decoding. *The EMBO Journal*, 30(8):1497–1507, mar 2011.

[80] Elmar Behrmann, Justus Loerke, Tatyana V. Budkevich, Kaori Yamamoto, Andrea Schmidt, Pawel A. Penczek, Matthijn R. Vos, Jörg Bürger, Thorsten Mielke, Patrick Scheerer, and Christian M.T. Spahn. Structural Snapshots of Actively Translating Human Ribosomes. 161(4):845–857, may 2015.

[81] Cristina Maracci, Frank Peske, Ev Dannies, Corinna Pohl, and Marina V. Rodnina. Ribosome-induced tuning of GTP hydrolysis by a translational GTPase. 111(40):14418–14423, sep 2014.

[82] Tina Daviter, Hans-Joachim Wieden, and Marina V. Rodnina. Essential role of histidine 84 in elongation factor tu for the chemical step of GTP hydrolysis on the ribosome. *Journal of Molecular Biology*, 332(3):689–699, sep 2003.

[83] A. Aleksandrov and M. Field. Mechanism of activation of elongation factor tu by ribosome: Catalytic histidine activates GTP by protonation. *RNA*, 19(9):1218–1225, jul 2013.

[84] Johan Åqvist and Shina C.L. Kamerlin. Exceptionally large entropy contributions enable the high rates of GTP hydrolysis on the ribosome. *Scientific Reports*, 5(1), oct 2015.

[85] Alexandra T.P. Carvalho, Klaudia Szeler, Konstantinos Vavitsas, Johan Åqvist, and Shina C.L. Kamerlin. Modeling the mechanisms of biological GTP hydrolysis. *Archives of Biochemistry and Biophysics*, 582:80–90, sep 2015.

[86] Jeffrey K. Noel and Paul C. Whitford. How EF-tu can contribute to efficient proofreading of aa-tRNA by the ribosome. *Nature Comm.*, 7:13314, oct 2016.

[87] Marina V. Rodnina, Niels Fischer, Cristina Maracci, and Holger Stark. Ribosome dynamics during decoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1716):20160182, jan 2017.

[88] Neş'e Bilgin, Leif A. Kirsebom, Måns Ehrenberg, and Charles G. Kurland. Mutations in ribosomal proteins l7/l12 perturb EF-g and EF-tu functions. *Biochimie*, 70(5):611–618, may 1988.

[89] Wei Liu, Chunlai Chen, Darius Kavaliauskas, Charlotte R. Knudsen, Yale E. Goldman, and Barry S. Cooperman. EF-tu dynamics during pre-translocation complex formation: EF-tu·GDP exits the ribosome via two different pathways. 43(19):9519–9528, sep 2015.

[90] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. 30(16):2785–2791, Dec 2009.

[91] Andrea Parmeggiani, Ivo M. Krab, Sumio Okamura, Rikke C. Nielsen, Jens Nyborg, and Poul Nissen. Structural basis of the action of pulvomycin and ge2270 a on elongation factor tu,. *Biochemistry*, 45(22):6846–6857, 2006. PMID: 16734421.

[92] Matthew J. LaMarche, Jennifer A. Leeds, Adam Amaral, Jason T. Brewer, Simon M. Bushell, Gejing Deng, Janetta M. Dewhurst, Jian Ding, JoAnne Dzink-Fox, Gabriel Gamber, Akash Jain, Kwangho Lee, Lac Lee, Troy Lister, David McKenney, Steve Mullin, Colin Osborne, Deborah Palestrant, Michael A. Patane, Elin M. Rann, Meena Sachdeva, Jian Shao, Stacey Tiamfook, Anna Trzasko, Lewis Whitehead, Aregahegn Yifru, Donghui Yu, Wanlin Yan, and Qingming Zhu. Discovery of LFF571: An investigational agent for Clostridium difficile Infection. 55(5):2376–2387, mar 2012.

[93] D. Branduardi, F.L. Gervasio, and M. Parrinello. From A to B in free energy space. 126:05103, 2007.

[94] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio

Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. 180(10):1961 – 1972, 2009.

[95] Barry Israelewitz, Mu Gao, and Klaus Schulten. Steered molecular dynamics and mechanical functions of proteins. 11(2):224–230, 2001.

[96] Alessio Lodola, Davide Branduardi, Marco De Vivo, Luigi Capoferri, Marco Mor, Daniele Piomelli, and Andrea Cavalli. A catalytic mechanism for cysteine n-terminal nucleophile hydrolases, as revealed by free energy simulations. *PLoS ONE*, 7(2):e32397, feb 2012.

[97] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. 100:020603, Jan 2008.

[98] Alan Grossfield. WHAM: the weighted histogram analysis method. 2013.

[99] Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. 31(2):455–461, 2010.

[100] A. VanWart, J. Eargle, Z. Luthey-Schulten, and R. Amaro. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. 8:2949–2961, 2012.

[101] Kazutaka Katoh and Daron M. Standley. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, 32(13):1933–1942, Feb 2016.

[102] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. Weblogo: A sequence logo generator. *Genome Research*.

[103] Michael Jensen, Robbert H. Cool, Kim K. Mortensen, Brian F. C. Clark, and Andrea Parmeggiani. Structure-function relationships of elongation factor tu. *European Journal of Biochemistry*, 182(2):247–255, 1989.

[104] Kirill B. Gromadski, Hans-Joachim Wieden, and Marina V. Rodnina. Kinetic mechanism of elongation factor ts-catalyzed nucleotide exchange in elongation factor tu†. *Biochemistry*, 41(1):162–169, jan 2002.

[105] J. M. Schrader, S. J. Chapman, and O. C. Uhlenbeck. Tuning the affinity of aminoacyl-tRNA to elongation factor tu for optimal decoding. 108(13):5215–5220, Mar 2011.

[106] Gemma Atkinson. The evolutionary and functional diversity of classical and lesser-known cytoplasmic and organellar translational GTPases across the tree of life. *BMC Genomics*, 16(1):78, 2015.

[107] Alfred Wittinghofer and Ingrid R. Vetter. Structure-function relationships of the g domain, a canonical switch motif. *Annual Review of Biochemistry*, 80(1):943–971, jul 2011.

[108] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. 81(1):511–519, 1984.

[109] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, Mar 1985.

[110] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.

[111] Shuichi Miyamoto and Peter A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. 13(8):952–962, Oct 1992.

[112] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. 18(12):1463–1472, 1997.

[113] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, Feb 2013.

[114] Oliver F. Lange and Helmut Grubmuller. Generalized correlation for biomolecular dynamics. 62(4):1053–1061, 2006.

[115] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[116] Michael Waskom, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Alistair Miles, Tom Augspurger, Tal

Yarkoni, Tobias Megies, Luis Pedro Coelho, Daniel Wehner, Cynddl, Erik Ziegler, Diego, Yury V. Zaytsev, Travis Hoppe, Skipper Seabold, Phillip Cloud, Miikka Koskinen, Kyle Meyer, Adel Qalieh, and Dan Allan. Seaborn: v0.5.0 (November 2014), 2014.

[117] Jos M. Raaijmakers, Timothy C. Paulitz, Christian Steinberg, Claude Alabouvette, and Yvan Moënne-Loccoz. The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms. *Plant and Soil*, 321(1-2):341–361, feb 2008.

[118] Carmen M. Herrera, Maria D. Koutsoudis, Xiaolei Wang, and Susanne B. von Bodman. Pantoea stewartii subsp. stewartii exhibits surface motility, which is a critical aspect of stewart's wilt disease development on maize. *Molecular Plant-Microbe Interactions*, 21(10):1359–1370, oct 2008.

[119] Matthieu Barret, John P. Morrissey, and Fergal O'Gara. Functional genomics analysis of plant growth-promoting rhizobacterial traits involved in rhizosphere competence. *Biology and Fertility of Soils*, 47(7):729–743, jul 2011.

[120] Anelise Beneduzi, Adriana Ambrosini, and Luciana M.P. Passaglia. Plant growth-promoting rhizobacteria (PGPR): their potential as antagonists and biocontrol agents. *Genetics and Molecular Biology*, 35(4):1044–1051, 2012.

[121] Stijn Spaepen, Jos Vanderleyden, and Roseline Remans. Indole-3-acetic acid in microbial and microorganism-plant signaling. *FEMS Microbiology Reviews*, 31(4):425–448, jul 2007.

[122] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews. Microbiology*, 10(4):291–305, 2012.

[123] Caroline B Milne, Pan-Jun Kim, James A Eddy, and Nathan D Price. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnology journal*, 4(12):1653–1670, 2009.

[124] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and

Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, nov 2016.

[125] J. M. Monk, P. Charusanti, R. K. Aziz, J. A. Lerman, N. Premyodhin, J. D. Orth, A. M. Feist, and B. O. Palsson. Genome-scale metabolic reconstructions of multiple escherichia coli strains highlight strain-specific adaptations to nutritional environments. 110(50):20338–20343, nov 2013.

[126] Edward J O'Brien and Bernhard O Palsson. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Current Opinion in Biotechnology*, 34:125–134, aug 2015.

[127] Adam M Feist and Bernhard Ø Palsson. The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. *Nature Biotechnology*, 26(6):659–667, jun 2008.

[128] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9):977–982, aug 2010.

[129] The doe systems biology knowledgebase: Microbial communities science domain.

[130] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011. *Molecular Systems Biology*, 7(1):535–535, apr 2014.

[131] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. 44(D1):D457–D462, 2016.

[132] S. Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, sep 1997.

[133] J.M. Kaper and H. Veldstra. On the metabolism of tryptophan by agrobacterium tumefaciens. *Biochimica et Biophysica Acta*, 30(2):401–420, nov 1958.

[134] JAMES H. M. HENDERSON. Biological activity of degradation products of indolepyruvic acid. *Nature*, 205(4972):702–703, feb 1965.

[135] Volker Magnus, Šumski Šimaga, Sonja Iskrić, and Sergije Kveder. Metabolism of tryptophan, indole-3-acetic acid, and related compounds in parasitic plants from the genus orobanche. *Plant Physiology*, 69(4):853–858, 1982.

[136] Piyush Labhsetwar, John Andrew Cole, Elijah Roberts, Nathan D. Price, and Zaida A. Luthey-Schulten. Heterogeneity in protein expression induces metabolic variability in a modeled escherichia coli population. 2013.

[137] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele, and D. Schomburg. BRENDA, the enzyme information system in 2011. 39:D670–D676, Nov 2010.

[138] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. 329(5991):533–538, jul 2010.

[139] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2016-08-23].

[140] Stefan van der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, 13(2):22–30, mar 2011.

[141] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Systems Biology*, 7(1):74, 2013.

[142] Fernando Perez and Brian E. Granger. IPython: A system for interactive scientific computing. *Comput. Sci. Eng.*, 9(3):21–29, 2007.

[143] Zachary A. King, Andreas Dräger, Ali Ebrahim, Nikolaus Sonnenschein, Nathan E. Lewis, and Bernhard O. Palsson. Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLOS Computational Biology*, 11(8):e1004321, aug 2015.

[144] Els Prinsen, Walter Van Dongen, Eddy L. Esmans, and Henri A. Van Onckelen. Hplc linked electrospray tandem mass spectrometry: A

rapid and reliable method to analyse indole-3-acetic acid metabolism in bacteria. *Journal of Mass Spectrometry*, 32(1):12–22, 1997.

[145] G.B. Kulkarni, A.S. Nayak, S.S. Sajjan, A. Oblesha, and T.B. Karegoudar. Indole-3-acetic acid biosynthetic pathway and aromatic amino acid aminotransferase activities inPantoea dispersastrain GPK. *Letters in Applied Microbiology*, 56(5):340–347, mar 2013.

[146] M Heldal, S Norland, and O Tumyr. X-ray microanalytic method for measurement of dry matter and elemental content of individual bacteria. *Applied and Environmental Microbiology*, 50(5):1251–1257, 1985.

[147] Ron Milo. What is the total number of protein molecules per cell volume? a call to rethink some published values. *BioEssays*, 35(12):1050–1055, sep 2013.

[148] Yi Tao, Jean-Luc Ferrer, Karin Ljung, Florence Pojer, Fangxin Hong, Jeff A. Long, Lin Li, Javier E. Moreno, Marianne E. Bowman, Lauren J. Ivans, Youfa Cheng, Jason Lim, Yunde Zhao, Carlos L. Ballaré, Göran Sandberg, Joseph P. Noel, and Joanne Chory. Rapid synthesis of auxin via a new tryptophan-dependent pathway is required for shade avoidance in plants. *Cell*, 133(1):164–176, apr 2008.